

COMPARISON OF NATURAL AND SYNTHETIC
LARGE INVERSELY SYMMETRIC SPECTRA

By

WILSON MARAL SILVA

A DISSERTATION PREPARED TO THE GRADUATE COUNCIL OF
THE UNIVERSITY OF FLORIDA
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1974

"ALL are here, but only the
immature and inhibited are."

WILLIAM SHAKESPEARE

To America's Most/less Knowledgeable

ACKNOWLEDGMENTS

I wish to acknowledge the assistance given to me by those who made possible the completion of this work.

First, I want to express gratitude to my advisor, Dr. Donald E. Childers, whose extraordinary devotion, professional competence, and sincere encouragement have made my task, in my estimation, most fruitful. I take pride in bearing testimony that my association with him will provide a lifelong sustenance and inspiration.

I also wish to acknowledge the available assistance given to me by the members of my committee, Drs. E. Fred Moore, Arnold Krige, Leon W. Smith, and Nathan E. Perry.

In addition, I am grateful to others who assisted me. Thanks are due to Mrs. Bernice Romano, for her assisting effort on my behalf. Her cooperation, in typing the manuscript, made possible the completion of the work on time. Appreciation is expressed to Dr. Beverly H. Willstrand, for her time and labor spent on the draftings, and to Dr. William A. Teet, for his guidance and aid in the procedural setup of the listener evaluation session.

Recognition is due to the University of Florida Board, Stephen Cooper, for granting my request and for the financial support. In particular, thanks are due to Prof

Refael Ariel Oren, Chancellor, for his continued encouragement, and for the faith that he has placed in me.

Finally, I wish to express my indebtedness to my parents, relatives, and friends, especially to Dr. Theodore Mordecai Jinkins, my former teacher and present colleague, whose encouragement and help have undoubtedly stimulated me in this work.

This research was supported in part by grant RR 11718 from the National Institute of Neurological and Communicative Disorders and Stroke of the National Institutes of Health.

TABLE OF CONTENTS

	<u>Page</u>
ACKNOWLEDGMENTS	iii
ABSTRACT	vii
CHAPTER	
I INTRODUCTION	1
Statement of the Problem Objectives and Rationale	1
Summary of Thesis	8
Chapter Summaries	12
II THE VOTAL AREA FUNCTION	13
Definition of Terms	14
Measurement of the Votol Area Waveform	18
Speech Synthesis based on Votol Area Waveform Formalism	24
Statement of the Research Problem	25
III HISTORICAL BACKGROUND ON METHODS OF SPEECH SYN- THESIS	27
Synthesis Procedures Selected	30
IV METHODS OF EVALUATION	31
Procedures for Synthesizing Speech	31
Synthetic Speech Evaluation	39
Procedures for Pathological Cases	40
Summary	46
V RESULTS	47
Synthetic Speech Results	47
Diagnostic Evaluation Results	47
Pathological Case Results	50

TABLE OF CONTENTS (continued)

CHAPTER	Page
VI DESCRIPTION	78
Global Area Remarks	80
Synthesis and Real Speech Remarks	82
Orthological Case Remarks	83
Concluding Remarks	84
APPENDIX	
A. SPEECH SYNTHESIS BASED ON VOWEL CORE MODELING	99
B. METHOD OF LINEAR PREDICTION IN SPEECH SYNTHESIS	108
C. PROGRAMS	117
Source Listings	117
Example	118
D. LISTENING EVALUATION FORM	119
E. DATA FOR RESULTS IN AND OF	124
REFERENCES	130
BIBLIOGRAPHICAL SHEET	134

Abstract of Dissertation Presented to the Graduate Council
of the University of Florida in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

COMPARISON OF NATURAL AND GLOTTAL
AREA WAVEFORM SYNTHETIC SPEECH

By

Karlann Rodal Parls

December, 1994

Chairman: Donald E. Childers
Major Department: Electrical Engineering

The objective of this research was to investigate the role of the vocal cord vibratory pattern in the production of natural voiced speech. The naturalness of synthetic speech, derived from glottal area waveform excitation, was evaluated and compared to real speech by conducting a casual listener session. Three tasks were established to determine whether or not the real and synthetic speech were indistinguishable, and to judge the quality of the synthetic speech.

The glottal area waveform, obtained from a pathologized subject, was also used to synthesize speech. The results are presented in this study.

The glottal area was measured from high-speed film obtained by photographing the vocal cords at an average speed of 4,000 frames/s. The measurement of the glottal area was automated, and the equipment included a Dene 1 minicomputer, a Tektronix 4625 CRT display terminal, a Graftek, and supporting software.

A semi-automated software system was developed to synthesize speech from the glottal area waveform

The vocal tract was modeled by a digital filter, whose parameters were found from the method of linear prediction. The steps to derive the model were as follows. The sound phonated during the filming session was recorded on magnetic tape, sampled at 12 KHz and stored on computer disk. A record of three pitch periods of speech was then used to determine the linear prediction all-pole filter model. The formants were extracted from the linear prediction model by factoring it and retaining the dominant poles. Formant bandwidths were changed according to the current literature, since it is believed that the linear prediction method does not give correct bandwidth values. The factored filter was then reassembled and the resultant expression employed as the vocal tract model.

Synthetic speech was obtained by parabolically warping the glottal area measured from high-speed films to an effective 12 KHz, parabolically extending one glottal pitch period, incorporating the filter and shifter procedures derived from the original speech into the resultant glottal area waveform and exciting the vocal tract model with this function. The effect of radiation at the mouth was also considered.

Three glottal area waveforms, obtained from normal male subjects, were used to synthesize voiced speech. For the listener evaluation, a magnetic tape was prepared containing one-second speech segments of both the synthetic and real speech obtained during the filming session. The listener

evaluation tests showed that the synthetic speech differed substantially from its natural counterpart, that is, they were distinguishable. Furthermore, the judges demonstrated a strong preference for the real speech by indicating that it sounded more natural than the synthetic speech. Based on these findings and on current research being done in speech sound modeling, we concluded that the sine waveform was not adequate for a complete description of speech sounds. The reasons for this are discussed at length.

Results from the pathological case indicated that glottal and alveolar, when introduced into the glottal waveform, accounted for the harsh sound quality present in the real speech.

In general, considerable trial and error work had to be done in the analysis of the pathological case due to the lack of available information concerning certain speech parameters.

CHAPTER 1 INTRODUCTION

Statement of the Problem Objectives and Background

At the turn of this half century a surge of interest developed in the scientific community in the area of speech analysis and synthesis. Both the engineering and medical communities took an active interest in this field. On the technological side, engineers looked for more efficient ways to transmit the speech signal at lower bit rates while maintaining its intelligibility and, whenever possible, its naturalness. On the medical side, physiologists began to investigate how our vocal apparatus worked in the hope that this would aid the diagnosis of voice disorders.

Recently, we have witnessed a combination of efforts in speech analysis and synthesis on the part of both of these scientific communities. There are many reasons for this. Most speech synthesizers are modeled after our vocal mechanisms, thus requiring a knowledge of its physiology. It is exceedingly difficult to study the human vocal mechanism due to its inaccessibility, thus physiologists are in need of models that can simulate vocal behavior to aid their study of our vocal mechanism. Speech pathologists are in need of automated procedures to produce speech data which,

without such mechanical variations, requires large amounts of human time and effort. This data processing can be done effectively and quickly with the aid of high-speed digital computers. A flow diagram showing the interactions between the two the engineering and medical fields in speech-related areas is given in Figure 1.1.

As we just mentioned, modeling of the human vocal mechanism plays an important role in most speech synthesis systems. Relatively little is known about our vocal system also because the problems involved in studying its behavior are multiple. Hence, many models actually simulate the vocal mechanism's overall characteristics without taking into account its detailed structure. This usually leads to intelligible synthetic speech which lacks the natural quality normally found in human speech. It thus seems advantageous to study in more depth those parameters of the vocal apparatus that contribute to the naturalness of speech.

Figure 1.2 illustrates a simplified diagram of the human speech production mechanism. The vocal cords can vibrate and are responsible for the production of voiced speech. The surface between the vocal cords is called the glottis and as they open and close, that is, vibrate, they describe an area. The volume velocity of air (rate of air flow) that passes through the glottis is modified by the pharynx, mouth and nasal cavities to produce speech. If the speech is produced as a result of the vibratory action of the vocal cords it is voiced, otherwise, it is unvoiced. Thus, all voiced speech sounds originate at the vocal cords

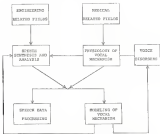


Figure 1.1. Flow Diagram shows the interaction between the engineering and medical fields in speech related areas.

The objective of the research to be described here was to investigate the role of the vocal cord vibratory pattern in the production of voiced speech. It was to be determined if the vocal cord vibratory pattern contained parameters sufficient for the production of natural speech. In order to accomplish this task a digital filter model of the vocal tract whose parameters were obtained by the linear prediction method [2] was computer-implemented using the digital area method, which was derived from high-speed films of

tract filter model to synthesized speech. The area waveform, obtained from a pathological case, was also used to synthesize speech.

Figure 2,3 summarizes the steps that were followed to achieve the objectives of this research.

Summary of Results

The results obtained were dependent upon a number of techniques developed in an attempt to optimize the quality of the synthetic speech. These techniques are also mentioned here so that the reader can gain greater insight to the importance of the results obtained.

It was desirable to obtain synthetic speech at a sampling rate of 10 KHz since it is well known that the bandwidth of speech is of the order of 5 KHz [1]. The glottal area waveform, measured from high-speed films of the vocal cords, was sampled by the camera at approximately 4,400 frames/s. A parabolic interpolation scheme was implemented, among others, to artificially increase the sampling rate of the area waveform to 10 KHz. Results showed that the quality of the synthetic speech was affected by the interpolation of the glottal area waveform since every frequency component in the synthetic speech above 2.2 KHz was the result of the interpolation process.

For the synthesis, only a single pitch period of the glottal waveform was used as the input to the vocal tract filter. There were two reasons for doing this. First, when



Figure 1-7 Flow diagram illustrating the steps leading to the research objectives.

speed variations made it impractical to use long record lengths of the glottal area and instead, speaker movements of the nose waveform produced noticeable differences in our remove pitch periods. The single pitch period was then periodically extended at the input to the filter to produce approximately one second of speech. The speech obtained sounded monotonous due to the lack of jitter and shimmer in successive pitch periods. The original speech signal, recorded during the filtering session, was then analysed to extract its jitter and shimmer parameters. These were then incorporated into the synthetic speech to give it more natural qualities.

The subjects, while being filmed, were instructed to attempt to phonate an [i] sound. This had the effect of lifting the epiglottis, which normally blocks the anterior portion of the vocal cavity from view. However, the sound actually emitted was an [ae] sound. It was difficult to obtain any other voiced sound because the filming process required the insertion of a laryngeal mirror into the subject's mouth. The subjects were thus limited to the production of simple voiced sounds or phonemes.

The vocal tract filter model also had to undergo some modifications. The method of linear prediction when applied to the speech signal models the vocal tract, the effects of glottal excitation, and the effects of radiation at the mouth and nostrils. The resultant model is an all-pole filter. For speech sampled at 12 KHz the order of the filter is usually between 12 and 14. In the range of 1 KHz the vocal tract approximates 4 or 5 formants. This means that the vocal tract filter model should have 4 or 5 complex conjugate poles. The other poles in the linear prediction model account for the overall effects of radiation and glottal excitation. The linear prediction filter model was then factored and only the poles describing the vocal tract were kept. The formant bandwidths were then allowed to agree with the current literature [1], and the poles multiplied to form the new filter.

The speech obtained, after exciting the vocal tract filter model with the glottal area waveform, had to be

multiplied in the exponent by the factor $(1-\alpha)^{-1}$, as necessary for evaluation at the roots; experimental results showed that the optimum synthetic speech was produced for $\alpha(1-\alpha)^{-1} = .6(1-.75)$.

Figure 1-4 shows the sequence of the different tasks which were developed in this research.

Three glottal wave waveforms, obtained from normal male subjects, were selected to synthesize voiced speech. Three tasks were developed for the listener evaluation of the synthetic speech. The different tasks permitted the listeners to compare the real speech to its synthetic counterpart, and to judge the quality of the synthesized speech in terms of its naturalness. The first task consisted of segments of both natural and synthesized speech; the listener's task was to decide, after hearing each segment, if it sounded natural. In the second task the listeners were presented with pairs of speech segments; each pair contained a natural segment along with its synthetic counterpart. The listeners then selected the segment in each pair which sounded most like natural speech. The last task again consisted of pairs of speech, but this time the listeners were to determine if the two segments in each pair sounded the same.

In general, the results from the listener evaluation tasks showed that the synthetic speech differed substantially from its natural counterpart, that is, they were distinguishable. Furthermore, the judges demonstrated a strong prefer-



case for the real speech by indicating that it sounded more natural than the synthetic speech. We have since determined the major reasons for these results. These are

- i) photographic difficulties
 - a) camera speed variations
 - b) insufficient camera speed
 - c) offset of temporal mirror
- ii) difficulty in measuring the area waveform from film.
- iii) format bandwidth restrictions,
- iv) assumption that source-trust are linearly separable and,
- v) the area waveform may not contain the parameters sufficient for the production of natural speech,

which we discuss in more detail in Chapters II and VI.

Results from the pathological case indicated that jitter and shimmer, when introduced into the glottal waveform, approximated for the harsh sound quality present in the real speech. No assessment concerning the naturalness of the synthetic speech could be made due to its pathological nature.

The major contributions of this research are

- i) an identification of the details of the vocal cord area waveform needed for the production of natural speech
 - a) jitter
 - b) shimmer
 - c) effects of the interpolation process
- ii) experimental modifications of the vocal tract filter to optimize the synthetic speech
 - a) formant extraction
 - b) formant bandwidth modification
 - c) the incorporation of radiation effects
- iii) the first attempt to produce synthetic speech by directly using the glottal area waveform to create a digital filter model of the vocal tract, and
- iv) a direct assessment of the adequacy of using the glottal area waveform to synthesize natural speech.

Chapter Summaries

The physiology of the vocal mechanism played an important role for this study. Hence, many terms, normally not encountered in the engineering literature are employed in this work. Chapter II, based on the glottal area function, first describes the vocal mechanism and gives an explanation of the technique used in this research to measure the glottal area waveform. The chapter concludes with two sections, one covering the Filter model used in synthetic speech and the other describing the synthesis problem of this research. Chapter III is a historical overview on the methods of speech synthesis. The purpose of this chapter is to give a narrative on the development of speech synthesis and at the same time to review the literature pertinent to this field. The details for synthesizing speech are discussed in Chapter IV. Some of the problems encountered in this research are also described and analyzed. Chapter V presents the results obtained from this investigation. The first section of this chapter describes the resultant synthetic speech of one subject. Graphs are plotted so that the reader can compare the read with the synthetic speech. The next section tabulates the results from the listener evaluation of the natural speech and the third section presents the results obtained from the psychological test. Chapter VI summarizes the results obtained from this research.

Five appendices have been included. Appendix A gives a review of speech synthesis based on vocal cord modeling.

The purpose for including this as the dissertation is that it has played a relevant part in the analysis of the results of this research. Appendix A reviews the method of linear prediction, which was used for formant extraction. Appendices C and E give, respectively, a listing of the main computer programs and a listing of the listener evaluation form. Appendix F gives the synthetic speech data for two of the normal subjects.

CHAPTER II

THE GLOTTAL AREA FUNCTION

The physiology of the vocal mechanism plays an important role in the design and implementation of speech synthesizers and cannot be overemphasized. Most speech synthesizers, in one way or another, model our vocal apparatus. There are basically two types of modeling devices: one class views the vocal apparatus as an input-output device. This type of synthesizer models the overall characteristics of the vocal mechanism but does not take into account structural details. A second class of synthesizers, called articulatory or transmission line synthesizers, models the vocal mechanism by dividing it into sections. Each section is then modeled by an electrical network.

The model of the vocal mechanism employed in this research falls into the category of input-output devices. The excitation function for the model is the glottal area waveform, obtained from high-speed photography of the vocal cords. We shall briefly describe the physiological aspects of the vocal mechanism that played an important role in this study. This chapter will, therefore, describe the human voice production mechanism and define some basic terminology. The measurement and use of the glottal area waveform in speech synthesis are

then discussed. The chapter concludes with a summarizing of the research problem.

Definition of Terms

A schematic diagram of the human speech production mechanism appears in Figure 1.1. Speech is produced by the driving force of the lungs expelling air through the bronchi, trachea, past the vocal cords, larynx, pharynx, and through the mouth and/or nasal cavity. The vocal tract is defined to be the passageway starting at the vocal cords and terminating at the lips and nose. The different sounds are produced according to the physical configuration of the vocal tract.

Speech sounds can be classified into two general sub-headings, voiced or unvoiced. Voiced sounds are always produced by the vibratory action of the vocal cords. Thus, all voiced sounds originate at the vocal cords. Voiced sounds are called nasals if they are produced through the nasal cavity (for example [n]). Unvoiced sounds are produced with the vocal cords spread apart by allowing either of two things to occur: on the first, the air stream passes through a narrow constriction in the vocal tract, in which case turbulent flow is produced. In the second case, total closure occurs at some point in the vocal tract followed by a pressure buildup and rapidly released it. Examples of unvoiced sounds include all of the voiceless. Examples of unvoiced sounds are s (for the first nasal) and z (for the second nasal).

In general, the transmission properties of the vocal tract are characterized by both resonances (peaks) and antiresonances (zeros). These resonances, which appear as spectral peaks in the frequency domain, are called *formants*. For unvoiced sounds the vocal tract's transfer function consists only of poles [1]. To obtain an idea of the location of the formant frequencies consider the unconstricted vocal tract of a male adult. For a first-order approximation we can model the vocal tract by a uniform cylindrical tube closed at one end and open at the other. The formants of this vocal tract model occur at the natural frequencies f_n of the tube. In equation form $f_n = (2n-1)c/4L$ where c is the velocity of sound in air (34,000 cm/s) and L is the length of the tube [1, p. 208]. For the normal adult male $L \approx 17$ cm, and we have $f_n = 2n-1150$ Hz. In the range from 0 to 5 kHz we then have five equally spaced formants. For the vocalized case, however, the vocal tract is a nonuniform anatomical tube and the formant locations will depend upon the particular configuration of the vocal tract.

The vocal cords are also referred to as the *vocal folds* and the crevice between them is called the *glottis*. As the vocal cords vibrate they describe an area, A . This area is variously referred to as the *glottal area waveform*, *glottal area function*, or simply *glottal area* (the term *vocal cord* is sometimes substituted in place of the word *glottal*). For voiced sounds, the glottal area is a function of time, that is, $A = A(t)$. The wave $A(t)$ is periodic in nature and quasi-

periodic, that is, the larynx with periodic leads to periodicity with occasional variations in amplitude and temporal fluctuations, also known as shimmer and jitter, respectively. The closed phase and open period of one glottal cycle (pitch period) are defined to be, respectively, the interval in time for which $k(t) = 0$ and $k(t) \neq 0$. The open period divided by the pitch period is called the open quotient or duty cycle. A typical glottal area waveform, measured from high-speed films of the vocal cords, is illustrated in Figure 1.1.

During the production of voiced sounds puffs of air are generated at the glottis as it opens and closes. The air puffs pass through the vocal tract, the configuration of which determines the particular sound emitted. In order to characterize the flow of air through the vocal tract we define volume velocity as the volume of air per unit time. If $w(t)$ is the particle velocity at the glottis, we have volume velocity, $U(t) = w(t)A(t)$. From this equation we see that the glottal volume velocity and area waveforms are related by the time-varying factor $w(t)$.

Volume velocity at the glottis is also referred to in the literature as glottal flow. In this work, the term volume velocity will always be taken to mean glottal flow.

The vibratory action of the vocal cords is caused by a combination of subglottal air pressure and the Bernoulli effect. Assume that the vocal cords are initially open. As air is expelled from the lungs it passes through the

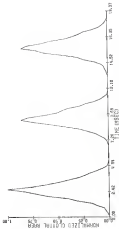


Figure 3.1: Normalized output power measured from high-speed films.

glottis, causing a drop in pressure. Since the walls of the vocal cords are flexible, they are brought closer together until full closure is achieved, at which time the Bernoulli effect ceases and the subglottal pressure increases steadily forcing the cords apart. The cycle is thus repeated in- during continuous oscillations at the vocal cords.

Measurement of the Glottal Area Waveform

There are various ways to measure the glottal area waveform. These methods include motion picture photography, cinematography, fluor optics and radiographic techniques. However, the method that allows a detailed examination of vocal cord movement is that of high-speed photography [8]. This photography is currently being carried out at the University of Florida's Department of Speech, Gainesville, Florida. The glottal area waveforms used in this research were obtained from high-speed films of the vocal cords. Hence, we will limit ourselves to discussing this technique. For a review on the previously mentioned techniques the reader is referred to [4-7].

The advantages of high-speed photography of the vocal cords are many. Parameters, such as vocal cord length, width and area can be obtained and at the same time a detailed examination of the glottal vibratory pattern is possible. The filming of a patient's vocal cords allows the speech pathologist to have a visual view of cord movement and the pathologist can thus make a direct assessment of vibratory pattern for clinical diagnosis.

There are, however, a number of difficulties involved in the filming of the vocal cords. A laryngeal mirror is inserted into the subject's mouth to reflect the vocal cord loop onto the camera lens. The subject must also hold his tongue in a forward position while being photographed to avoid getting the tongue between the mirror and the camera lens. Many subjects cannot tolerate the laryngeal mirror and most experience some discomfort. As a consequence the subject can only produce a limited number of voiced sounds. The most common sound is [ae].

For this study, the sound emitted from the subject while being photographed was simultaneously recorded on magnetic tape. It was also possible to synchronize the sound source to the film, frame by frame, by having a small neon light flash out a frame at the same time that a released, held burst to the subject's throat, was activated to produce a sharp sound picked up by the microphone. However, the camera noise was extremely high and this experiment could not be used. A sample, recorded just prior to the camera being turned on, was used in this study to avoid the noise.

A sampling rate problem arose as a direct result of the variable camera speed and the fact that the sampling rate of the area scanned by the camera was not equal to that of the speech sampling rate (16 kHz). Figure 2-2 shows the operating characteristics of the 16 mm Philips camera used to photograph the vocal cords. At no point is the camera speed constant and the highest attainable camera speed is

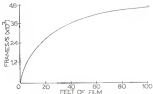


Figure 3.2 operating characteristic of Fostex high-speed camera.

approximately 4,800 frames/s. The film segment that was picked for analysis was between 70 and 48 feet of film. This interval was where the minimum camera speed fluctuations occurred. The camera speed in this interval was approximately 4,400 frames/s. After examining several interpolation schemes the Gregory-Newton difference interpolation formula (8) was used to increase the sampled rate of the glottal area waveform to an effective rate of 18 kHz.

Usually, the epiglottis covers a small portion of the vocal cords anteriorly and the arytenoid cartilages also tend to do the same, but posteriorly. Thus, the films may not show a full view of the vocal cords. This, and the fact

that focus on the vocal cords sometimes blocks the glottal opening, introduced some difficulties in the measurement of the area function from the film.

The usual procedure for measuring the glottal area from high-speed films is long and tedious. The glottal opening, first photographed on high-speed film, is traced on paper and from this the area is measured manually by using a mechanical instrument called a planimeter. Assuming that the person using the planimeter is highly trained, the only limitation of the area can be done at the rate of 18 frames/hour. Recent techniques, developed at the University of Florida's Department of Electrical Engineering, have improved the efficiency of the area measurement. Here, a mini-computer and other hardware are used to measure the glottal area by computerized means. Rates of 120 frames/hour are now possible. This is slightly over an order of magnitude greater than the planimeter method, making larger data analysis more feasible.

The equipment, illustrated in Figure 2.1, includes a Room 2 minicomputer, a Tektronix 4010 CRT display terminal, a Graf/pen, and supporting software. The Graf/pen consists of a stylus and two linear 15" microphones mounted on a glass sample tablet. The microphones and the stylus are connected to a control unit which is interfaced with the computer. The two microphones serve as the x and y coordinates to the stylus, which generates sound impulses whenever pressure is exerted. A 14 cm wide single-frame projector

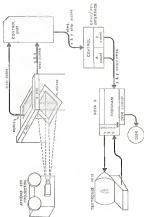


Figure 2.5 Block diagram of the data acquisition system (after RTI)

projects the film onto a piece of translucent paper placed on the acrylic surface. The operator traces the glottal opening using the Hoffman stylus. The sound impulses emitted by the stylus are picked up by the x-y microphones and the glottal outline is simultaneously displayed on the terminal. If the operator is satisfied with the outline he presses any key, except zero, on the display terminal. The area for the displayed frame is calculated and stored on disk file. The operator then advances the projector to the next frame and repeats the procedure. If the operator is not satisfied with the displayed glottal outline he presses the zero key on the terminal. This tells the computer to reject the frame and start over again. Once on disk file the data can be processed to extract parameters such as jitter, shimmer, open quotient, closed phase, etc. The glottal area waveform illustrated in Figure 3.1 was obtained by using this technique.

The measurement of the glottal area waveform, using the hardware described above, still requires a significant amount of operator intervention. The anterior and posterior portions of the vocal cords are usually covered by the epiglottis and the arytenoid cartilages, respectively. The operator can either omit these portions or he can attempt to extrapolate the glottal opening.

Any masses present on the vocal cords will also interfere with the glottal opening. At times the film is not clear and it is somewhat difficult for the operator to locate

the edges of the vocal cords. All of these factors can produce a significant amount of error in the shape of the area waveform. Present research being conducted at the University of Florida's Department of Electrical Engineering and Speech is aimed at overcoming these difficulties.

Speech Synthesis Based on Global Area Waveform Excitation

Investigations have already been performed in modeling the vocal tract and synthesizing speech by exciting the model with various waveforms. Rosenberg [8] and Solina [10] have both attempted to synthesize speech from vocal tract models by driving them with different glottal waveforms. Trautman, Shumaker and palgowski waveforms were used. Based on listener evaluations both of these investigations reported that the synthesized speech obtained had lost its naturalness as compared to real speech. However, the glottal area waveform was not used. In fact, the current literature does not seem to present any information on synthesized speech derived from glottal area waveform excitation. There are various reasons for this. Most researchers involved in speech synthesis are generally interested in finding efficient ways to synthesize speech. Their main goal is an overall representation of the speech signal and not a detailed one. Specialized equipment is needed to measure the vocal cord vibratory pattern. Currently, there are not more than 5 places in the United States involved actively in vocal cord high-speed photography [11].

Recently, vocal cord models have been developed by Flanagan *et al.*, [14-17] (see Appendix A). The models are based on physiological constraints allowing one to make analogies between the human vocal mechanism and the model. The vocal cord models make use of the glottal area waveform, derived from a system of differential equations, to find the volume velocity waveform. This allowed us to note the fundamental differences between the glottal area waveform and volume velocity. The results from Flanagan's vocal cord models were thus used in this study to help explain some of the results that were obtained in this research.

Statement of the Research Finding

It has been reported by van den Burg *et al.*, [8] that the volume velocity at the vocal cords is directly proportional to glottal area for large openings and to the cubic power of area for small glottal openings. In their one and two mass model of the vocal cords Flanagan *et al.*, [14-17] obtained noticeable detailed differences between the volume velocity and the glottal area waveforms. The area waveform was generally more smooth and symmetric than volume velocity. Source-filter interaction played a significant role in the shape of the volume velocity waveform and had little or no effect on the overall shape of the area waveform. Source-filter interaction was particularly evidenced by temporal detail on the volume velocity waveform due to the first formant of the vocal tract.

We see that there are subtle but significant differences between the volume velocity and glottal area waveforms. How relevant are these differences in regard to the audible speech signal? The purpose of this research was to seek an answer to this question. The method outlined is aimed at this question and to synthesize speech using the glottal area waveform, measured from high-speed films of the vocal cords, as the input to a vocal tract model. The formant frequencies for the vocal tract model were fixed by the linear prediction method (see Chapter III and Appendix B for details on this method). The synthetic speech and the original speech signal, obtained during the filming session, were used for listener evaluation tests. The listeners' tasks were to determine if the synthetic speech sounded natural and to compare the synthetic speech with the original speech.

The main goals of this research were,

- (1) to set up a synthetic procedure to synthesize speech from the glottal area waveform,
- (2) to compare glottal area waveforms synthetic speech with real speech,
- (3) to draw conclusions, based on the experimental results, relevant to the adequacy of synthesizing speech employing glottal area waveform excitation, and
- (4) based on (1)-(3), to recommend alternate pathways leading to a more useful analysis and interpretation of the glottal area waveform.

It was desired to implement a synthesis system, employing the glottal area waveform as the input to a vocal tract model, which required a minimal amount of operator intervention. Thus, an effort was made to reduce the amount of operator decisions with respect to the vocal tract (and other) parameters

CHAPTER 1.1

REVIEW(1) RECENT ADVANCES ON APPROACH ON SPEECH SYNTHESIS

An investigation of any subject is not complete without a knowledge of its historical aspects. The purpose of this chapter is to present an overview on the development of speech synthesis and at the same time to review the pertinent literature.

Speech synthesis can be traced as far back as the late 1700's. The objective in speech synthesis is to find a device that can imitate the human voice, while speech understanding systems hope to go beyond recognition. Multiple devices and algorithms have been employed in an effort to synthesize speech. Basically, the main goal in speech synthesis is to find an adequate model to represent the vocal tract. Devices have ranged from primitive mechanical synthesizers to sophisticated electronic ones and it is now possible, with the aid of high speed digital computers, to synthesize speech using programmed algorithms. To implement a mathematical model of the vocal tract using software has indeed been a significant breakthrough.

Von Kempelen's talking machine in the late 18th century was one of the first mechanical devices to successfully approximate speech [18-21]. An earlier device due to Erasmus

about 1850 was said to imitate the five vowels of the English language. It consisted of a set of five movable resonators, one for each vowel sound, where the form of each resonator resembled the human mouth cavity for the sound being emitted. Although Bratachevich's resonators did work, they were not as sophisticated in nature as von Selenius's machine, which could actually reproduce various speech utterances.

A modified version of von Selenius's speaking machine was built by Sir Charles Wheatstone in 1826 [13-15], and this in turn inspired Alexander Graham Bell during the late 1850's to construct a cast model of the human head which incorporated lips, tongue, palate, teeth, pharynx and velum made out of various different materials. Probably the most remarkable aspect was that it worked. Bell's device was able to produce voiced sounds and some simple utterances.

The years that followed brought no major contributions until Sir Richard Paget in the 1890's synthesized speech sounds using shaped resonant cavities excited in vibration by vibrating reeds. He showed that simple utterances could be produced by his innovations. He simply used his hands to form different cavity shapes. This resulted in the emission of various sounds, each one dependent on the particular cavity shape.

There were probably very many such devices, not even reported in the literature during the late years in the 19th century and the first few decades in this century. All of

These devices were mechanical in nature which made them somewhat cumbersome to handle. Nevertheless, as is still readily how many of these contraptions were even able to produce simple sentences considering their very limitations. Much like a musical instrument, the quality of speech sounds produced by these devices always depended on how good the operator was.

The transition between mechanical and electrical appliances is depicted out by two events. The production of the first electrical analog of the vocal organs was by Stewart [22] in 1912. Bell Telephone exhibited at the 1918 New York World's Fair the Voder (voice-operation demonstrator) invented by Homer Dudley [23-24], which could produce intelligible speech. Although Stewart's analog device could produce vocal-like sounds, the first electrical synthesizer to actually produce connected speech was the Voder [25]. The Voder consisted of six independent band-pass filters, which represented the vocal tract in terms of formant frequencies, and contained two different input sources. One was a white noise source, to simulate fricatives, and the other a periodic wave oscillator, to simulate voiced sounds. The end result was that there were about thirteen hand operated keys, one foot pedal for source pitch control and even a wrist bar for the selection of voiced or unvoiced sounds. Indeed, the operator had to be an expert. As a matter of fact the operator had to be trained for over a year before learning to properly use this machine. Thus, although electrical in

device, it still had the same major drawback as its mechanical predecessors, that is, it took an operator with considerable experience in work with the machine. However, the importance of the Voder and Stewart's electrical analog device are evident at this point. First, they filled the gap between mechanical and electrical modeling and second, their success proved that intelligible speech could be produced by electrical means, thus laying the foundations for present-day speech synthesis. The many devices that shortly followed, such as Dudley's Vocoder [24], came as a direct result of the Voder.

In 1938 Barrus [25], then at Bell Laboratories, designed a mechanical-electrical speech synthesizer. It was probably one of the first machines to synthesize speech from speech segments stored on magnetic tape. After a certain number of keys were struck, each one corresponding to a particular speech sound, the speech synthesizer would chain all of the sounds together. A time delay was incorporated near the output stage so that the segments would be somewhat continuously connected. Figure 3-1 shows the main idea behind this. The end result was what Barrus called *standardized speech* [25], that is, speech which has been synthesized from a connected sequence of bits of sound which are called "building blocks" of speech. Although still primitive in nature, his device set forth the initial stages for speech synthesis by this and at the same time made it clear that it was possible to fully automate the process of synthesizing speech.



Figure 3.1. Block diagram of speech synthesizer (after [24])

During the 1950's researchers started implementing electrical analog devices of the vocal tract. The main idea of these devices was to approximate the characteristics of the vocal tract by electrical networks supplemented by external control units to change its parameters. Two different types of devices for modeling the vocal organs developed. The first of them, called articulatory or transmission line analog synthesizers [1,21], view the vocal tract as a transmission line. The model consists of dividing the vocal tract into many sections and then designing a simple electrical network for each section. Each section simulates the distributed properties of a small part of the vocal tract. When put in cascade the end result is a direct representation of the vocal tract in terms of its transmission line properties. The first articulatory synthesizer is probably due to Ross [18]. Ross's device consisted of 25 unitary T-sections where each section represented a 4.5 cm length and 4 cm^2 of area of the vocal tract. A similar passive device, due to Stevens, Szwedski and Farn [29], made a major improvement in that their device provided adjustments to simulate a range of cross-sectional areas from 4.17 to 17 cm^2 . This device had shunt-line pi-sections, each representing 4.5 cm of the vocal tract.

The second group of models, called resonance, formant or formant-resonating synthesizers [2,31,36], view the vocal tract as an input-output device. The goal here is to model the device so that the transfer function will be a close correlate to the formant structure of the natural vocal tract. The earliest such device is due to Stewart in 1933 [32], but it was not until 1953 that the first model in synthesizer complete sentences was developed by Lawrence [33]. Currently, a multiplicity of such devices are under study and development [33-36].

At this point it is safe to say that the turning point of speech synthesis has been due to the development of the digital computer. Speech analysis and synthesis, whether in the time or spectral domain, can be done in a matter of seconds making it possible to perform operations in real time. Such new methods as speech synthesis by rule, by nonacoustic filtering, by analysis-by-synthesis, and by linear prediction have been developed primarily as a consequence of the feasibility to implement them by computer.

The first method just mentioned, speech synthesis by rule, originated from Harris' classical paper [37] dealing with what he called standardized speech. Basically, the goal is to derive a set of rules, store them, and then use them to operate on a set of discrete symbolic inputs such as phonetic symbols, phoneticisms, and printed text. Although the speech obtained by some investigators [38,39] through this method has been highly intelligible, the main

Oppenheim's low structural model as compared to real speech. Nevertheless, the main advantages are the low input information rate rate [46] and that it needs a set of codes at does not need any previous information, such as frequency band, from natural speech samples. Figure 1.2 shows a description of such a system. A good application of this type of synthesis is in the information answering services, whereby one picks up the telephone, dials a reference number and asks a question. The computer processes the (spoken) question and answers, all in a matter of seconds.

Homomorphic analysis-synthesis of speech was first presented in two published works [47-48] in 1968-69. Figure 1.3 illustrates the synthesizer's configuration. The low-time cepstral information (COT) is actually the windowed inverse Fourier transform of the log magnitude of the Fourier transform. Figure 1.4 shows that implementation in single block diagram form. Oppenheim reports [48] that this system can produce good high quality natural speech. We also see that homomorphic filtering is related to cepstral analysis.

The method of linear prediction and speech synthesis based on vocal tract modeling are described in Appendices B and C respectively. They are mentioned here only briefly to place them in perspective with the previous developments.

The linear prediction method was studied by Markel [49], Itakura and Saito [48], Atal and Rabiner [42], and a complete review was presented by Markel and Pold [52] and Markel and Gray [51].

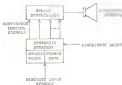


Figure 3.3 Rule of speech synthesis by rule based (SRI).

Linear prediction can actually be compared to the method of spectral analysis-by-synthesis first introduced at Bell Telephone Laboratories in 1942 [41]. The analysis-by-synthesis technique tries to find a spectral fit to the speech spectrum that consists of peaks and valleys. An error is then defined and the fit optimized by minimizing the error on the mean-squared error. The main difference between the two formulations is that in spectral analysis-by-synthesis the error minimization is done on the integrated logarithm of the ratio between the model and actual spectra whereas in linear prediction it is done on the integrated ratio of the two spectra. In general, the linear prediction error criterion leads to a better spectral envelope fit [44].

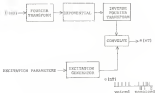


Figure 3.3 Representation of the synthesizer configuration after IMC2.

Essentially, the method of linear prediction models the voiced tract and the overall effects of radiation at the mouth and nostrils and of glottal excitation by an all-pole filter. This filter has the form

$$F(z) = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}} \quad (3.1)$$

where $z = e^{j\omega T}$, $s = \sigma + j\omega$ is a complex frequency variable, $T = 1/f_s$ is the sampling period and f_s is the sampling frequency.

We note that the linear prediction model is a unique one since it yields a single filter with all of the above mentioned properties. The fact that the filter does not



Figure 3.4 Representation of the analyzer configuration (after [11])

allow the user to vary the glottal pulse shape but limited only to a series of impulse functions so unlike the filter for voiced sounds is its chief failing for this research.

The pioneering work in vocal cord modeling is due chiefly to Flinker *et al.* [14-17]. Both one and two mass models of the vocal cords have been designed and implemented. In the one and two mass models each cord is represented, respectively, by one mass and two masses. The vocal cord models are coupled to a transmission line model of the vocal tract. The most important aspects of the vocal cord models are that their four input parameters are derived from physiological measurements. As a consequence, the effect of surgery-tract interaction is incorporated in the model.

Figure 3.5 summarizes the preceding discussion.

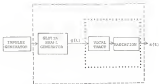


Figure 3-4 Model within dashed line is linear prediction model. The modified model used in this research is shown inside the dotted line.

SPARKS PROCESSING PRINCIPLE

The linear prediction method offers several advantages in speech synthesis: The model's filter coefficients are found in a quick and efficient manner leading to a recursive relationship. The model can then be implemented in the time domain. Furthermore, the simplicity of the model allows one to easily modify it.

This research used the linear prediction model for formant extraction. The model within the dashed line shown in Figure 3-4 is that derived by linear prediction. The model needed for this research appears within the dotted line in Figure 3-4. The procedure used for the filter modification is discussed in the next chapter.

CHAPTER 11

METHOD OF PROCEDURE

There were two major parts in this research. One was the development and implementation of the synthesis technique employing the glottal area waveforms to create a distorted filter model of the vocal tract. The other was the listener evaluation of the synthetic speech. This chapter discusses the detail of both of these procedures.

Procedure for Acoustic Area Search

Three normal male subjects were filmed for this research. They were asked to phonate a sustained [i] at their normal speaking frequency during the filming process (the sound that was actually recorded was an [ae] which is believed due to the laryngeal filter inserted into the subject's mouth). However, two of the subjects' vocal cords, while phonating at their normal speaking frequency, were not sufficiently exposed to allow an adequate view for photostereogram purposes (remember, the epiglottis covers the cords anteriorly, and the arytenoids posteriorly). The task-are was to increase their fundamental frequency and thus obtain a better view of the vocal cords at the cost of a decrease in temporal resolution per pitch period of the area waveform.

The timing of the vocal words, and the measurement and analysis of the glottal area waveform are described in the flow diagram of Figure 4.1. This sequence is outlined as follows. First, the vocal words were filmed using high-speed photography. The last 38 feet of a 4,000 frame 300-foot roll of film was selected for analysis since this was the region of optimum camera stability for short record lengths (see Figure 4.2). The area waveform of 300 consecutive frames was then obtained by a 100/psa interfaced to a Nova 2 minicomputer (see Chapter II for details). The 3-18 pitch periods obtained from the 300 frames were interpolated using the Gregory-Smith Difference Interpolation Formula [8] to artificially raise their sampled rate to an effective 10 kHz. The pitch periods were used separately to synthesize speech by periodically extending them. The jitter and release parameters, obtained from the original speech segment recorded during the filming sequence, was also incorporated into the synthetic speech. Each synthetic speech segment was then compared, by informal listening, to the natural speech. The segment that sounded most similar to the natural one was selected for the formal listener evaluation tests.

A first attempt at synthesizing speech using a period of 3-18 glottal pitch periods and periodically extending them proved to be fruitless. This was attributed to the following reason:

- (1) Camera speed variations were more apparent in long record lengths (>100 frames) as compared to shorter ones (<10 frames).



Figure 8-1 flow diagram describing the timing of the vocal mode, and the measurement and analysis of the glottal area waveform.

- (ii) Operator speech features: a) the measurement of the area waveforms due to bursts of the vocal cords, vocal tract, and partially nasopharyngeal cavity make the shape of each pitch period significantly different. It was also unclear if this difference was actually due to operator intervention or to glottal variations between successive pitch periods.
- (iii) A report of 8-12 pitch periods of glottal area data that contain sufficient information about pitch and timbre needed to characterize the speech signal.

Table 4.1 lists various glottal area waveform parameters for the 3 normal subjects photographed in this research.

Table 4.1. Glottal area waveform parameters averaged over 8-12 pitch periods. F_0 = fundamental frequency, Q_0 = open quotient, CP = open period, CP^* = closed period and CS = closure speed.

SUBJECT	F_0 (Hz)	Q_0	CP (ms)	CP^* (ms)	CS (cm/sec)
JB	129	.72	3.3	2.1	4926
SB	141	.73	4.1	1.9	4548
DB	146	.81	4.8	2.1	4910

In addition to the filming of the vocal cords the voiced sound phonated by the subject was recorded on magnetic tape. The sound recording served 3 purposes:

- i) It was used to derive the sound track filter coefficients.
- ii) The pitch and timbre parameters were obtained from this segment and incorporated into the synthesized speech.
- iii) This segment was used in the listener evaluation tests.

Specifically, the natural speech segment used should have been the one obtained simultaneously with the filming. However, camera noise was extremely high making it impossible to use this segment. The subjects were then asked to phonate before the initiation of the camera and this segment was then used for the purposes listed above. Care was taken to ensure that this speech sample had the same pitch period and intensity level as that during the filming.

The natural speech was lowpassed with a cutoff at 5 kHz, sampled at 18 kHz directly off the magnetic tape, and stored on a computer disk file for further processing. This allowed us to derive the vocal tract filter and to extract the jitter and shimmer parameters.

The method of linear prediction [7,8] was used to extract the vocal tract formant frequencies from the speech signal. (The linear prediction method is discussed in Appendix B). The filter, as derived by linear prediction, could not be used directly since the filter models adequately the vocal tract and the effects of glottal excitation and of radiation at the mouth and nostrils.

The flow diagram in Figure 4-2 illustrates the method employed in this research to construct the filter model of the vocal tract. Initially, the input parameters needed were approximately 3 pitch periods of speech data and the order p of the linear prediction filter. The linear prediction filter was then derived and its denominator factored into p poles. The s -plane poles were transformed to

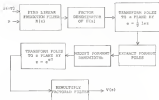


FIGURE 4-2 Flow diagram showing the method employed in this research to construct the filter model of the vocal tract. $V(s)$ is the vocal tract filter, and $x(nT)$ is the sampled version of $x(t)$.

the s plane by the relationship $s = \frac{1}{T} \ln z$, where T is the sampling period, and the conjugate pairs of poles corresponding to the forward frequencies were identified and kept. Since the sound phenomena were known a priori, it simplified the task of forward identification. The forward bandwidths were limited to conform to the current literature [7]. The forward poles were then transformed back to the s plane by $z = e^{sT}$ and the factored filter resampled. This linear filter served as the vocal tract model in this research. All of the computations to derive this filter were done by a FORTRAN program, named LPFORM1, implemented on a Univac 9

misconception. Appendix C gives a source listing followed by an example on how to use the program.

To illustrate the vocal tract modeling technique developed in this research, consider the plots shown in Figure 4.3. Figure 4.3a is the frequency response of a sixth order linear prediction filter derived from an 88 ms speech segment of subject 26. Essentially, the linear prediction filter models the envelope of the speech spectrum [44,47]. There is why it mimics the effects of glottal excitation and of radiation. This is apparent in Figure 4.3a from the additional resonance peaks present in the filter's frequency response. The first 4 formant frequencies were kept, reducing the order of the filter to $p = 4$, and the formant bandwidths were altered manually made smaller than those derived by the linear prediction model). The resultant frequency response of the modified filter is shown in Figure 4.3b.

The jitter and shimmer parameters introduced into the synthetic speech were obtained from the original speech segment. There are multiple ways to derive these parameters from the speech signal. No one method seems to be optimum. For example, the shimmer may be derived from the speech waveform by calculating the difference of the maximum amplitude in successive pitch cycles. The jitter may be found by first defining the period of each cycle C_j of the waveform to be the duration in time between the peak amplitudes of C_{j+1} and C_j . If the speech waveform consists of N cycles then

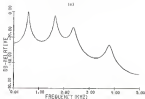
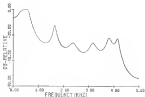


Figure 4.3 Frequency response of (a) linear prediction filter, $p = 25$ and (b) modified filter, $p = 5$.

$$P_{ij} = T_{ij} \delta_j - 1 \quad j = 1 \leq M+1$$

where T_j is the j th period and δ_j is the jitter parameter. However, for sampled speech these parameters, as defined above, were found to be inadequate when incorporated into the synthetic speech. The problem was that the sampled peak amplitude in each cycle was not a sufficiently good approximation to the actual peak amplitude, especially for the higher frequency speech waveforms. This introduced abrupt variations in the shinner and jitter between consecutive pitch cycles causing the synthetic speech to sound unnaturally harsh [10-13].

The jitter and shinner parameters employed in this research were found by a correlation analysis on the speech waveform, covering a random variation in these parameters from cycle to cycle. Figure 4.4 illustrates this method and the manner in which these parameters were incorporated into the synthetic speech. The shinner was derived by calculating the power in each speech cycle and multiplying this by the corresponding glottal cycle. Mathematically, if P_j is the power in the j th speech cycle and g_j is the j th glottal cycle, which is a function of time, we have

$$sint(t) = \sum_1^T P_j g_j$$

where $sint(t)$ is the glottal area waveform with shinner and T is the data sampling interval. Since for this research $g_j = g_k$ for all j, k , we have, letting $g_j = sint(t)$

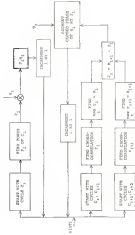


Figure 4.4 Flow diagram illustrating how the joint and slight parameters are extracted from the speech waveform and the technique used to incorporate them into the synthetic speech. Start is the speech waveform and T_L is the block period of the global area waveform.

$$Q(nT) = \left[\sum_{j=1}^N F_j \sin 2\pi n T_j \right] / T_j$$

where T_j is the pitch period of $q(nT)$.

The jitter parameter was derived by considering 3 successive speech cycles C_j , C_{j+1} and C_{j+2} . The cross-correlation functions r_j and r_{j+1} were found, respectively, for C_j and C_{j+2} , and for C_{j+1} and C_{j+2} . The ratios R_j and R_{j+1} were found for r_j and r_{j+1} , respectively, and the jitter was defined as

$$J_i = \max_k T_{i+1} - \max_k T_i = R_{i+1} - R_i \quad 1 \leq i \leq N-1$$

where N is the total number of cycles in the speech waveform. The jitter parameter was incorporated into the glottal waveform by repeating the glottal phase of each i th glottal cycle by J_i . A FORTRAN program called JITTER (due to Dr) A. Faig, Department of Electrical Engineering Faculty member, University of Florida), was modified and used to extract the jitter and shifter parameters from the speech waveform. The source listing appears in Appendix C.

To account for radiation at the mouth, the term $(1-s^{-1/2})$ was applied to the volume velocity wave from the vocal tract filter model as shown in Figure 4-3. In general, a value of $s = 1$ is used and the approximation $(1-s^{-1/2})$ to account for radiation is a good one for frequencies below 4 kHz [2,30]. The effect of $(1-s^{-1/2})$ is to differentiate the volume velocity wave $U_g(nT)$. Alternatively, the effect is that of high pass filtering where the higher frequency components are emphasized and the lower ones dampened.



Figure 4.4 Speech production model employed in this research. $V(s)$ is the vocal tract filter model and $u_g(t)$ is the volume velocity wave at the mouth corresponding to the input $g(t)$ glottal area waveform.

Experimental results showed that the value of $a = 1$ could not be used due to the undersampling of the glottal area waveform. Every frequency component beyond 2.25 kHz, for a source speed of 4.5 frames/s, was the result of the undersampling process. Suppressing these higher frequency components produced an unnatural quality in the synthesized speech. Informal listening to the synthetic speech for different values of a showed that a good value was $|a| \leq .75$.

The task of periodically extending the glottal area waveform, incorporating into it the shimmer and jitter parameters, and filtering it through the vocal tract model was done by a FORTRAN program, called *EXTENDED*. Another program, called *DIFFERENCE*, incorporated the effect of radiation. The final sequence of operations is illustrated in Figure 4.5. Appendix 2 gives a source listing of these two programs.

Synthetic Speech Evaluation

The final step consisted in conducting a formal listener evaluation of the speech synthesized from the 2 normal glottal

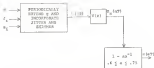


Figure 4-4 Signal sequence of operations for the synthesis of speech based on glottal area waveform excitation. The input parameters are jitter θ , shifter α , and one period T of glottal area waveform. $\theta_1(t)$ is the vocal tract filter excited by the glottal area waveform $\theta(t)$. $\theta_2(t)$ is the value modulo 2π of the sum of $\theta_1(t)$ and $\theta(t)$ is the resulting synthesized speech waveform.

area waveform. To this effect a separate tape containing both synthesized and natural speech segments was prepared. This section describes the organization of the tape and the manner in which the listener evaluation was conducted.

Three distinct tasks were set up for the listener evaluation session. In general, all 3 tasks consisted of segments of both the synthesized speech and its corresponding natural counterpart. The segments were all 1s in duration and their leading and trailing edges were multiplied by a linear ramp of duration 18 ms to avoid the sharp onset at the onset and offset. Before being recorded on tape all of the segments were low-pass filtered with a cutoff at 4 kHz.

In the first task the listeners were presented individual segments of synthetic and natural speech. All 8 segments of speech were used like 3 synthetic segments with their 3 corresponding natural counterparts and each segment was repeated 15 times for a total of 96 speech segments. The segments were distributed in random order and the spacing between each segment was 4 s. The 96 segments were divided into 3 groups, each containing 32 speech segments. The spacing between each group was 15 s. After hearing each segment the listeners responded whether they felt the segment sounded natural or unnatural. The purpose of this task was to judge the quality of the synthetic speech. The natural speech segments were also included to measure listener bias and to learn how natural was the natural speech.

The second task consisted of 16 pairs of speech segments. The spacing between pairs and between the elements within each pair were 1 s and .8 s, respectively. Each pair was composed of a synthetic speech segment along with its natural counterpart (the speech segment recorded during the training session). The 16 pairs were broken into 4 groups of 4 pairs and were randomly distributed. The spacing between the groups was 15 s. For this event the listener's task was to select the sample within each pair that sounded most like natural speech. Thus, the objective in this task was to compare the synthetic speech to the natural speech and to learn whether the synthetic speech had sufficient natural qualities to make it indistinguishable (in terms of naturalness) to natural speech.

The third task was similar in its structure to the speed task. The difference was that 24 pairs were used (12 groups of 12) and the listener's task was to reply whether the samples within each pair sounded the same or different.

The first task can be compared to a yes-no experiment and the second task to a two-alternative forced-choice (2AFC) procedure [10-12]. In a yes-no task the listener is presented a single stimulus and must make a decision on whether or not it has a certain given characteristic. In a 2AFC procedure the listener is presented with 2 samples and must select which of the 2 contains the desired characteristic.

The formal listener evaluation sessions were held in the same laboratory where the vocal cord films were done. A total of 4 female students, between the ages of 18-24, from speech-related sciences participated in the experiment. All of the students were at the advanced undergraduate level and all had some previous experience in clinical work dealing with voice disorders.

The listeners were seated in a quarter-circle and the speaker (professional Adult speaker) was placed in the center equidistant from them. Prior to the session a sound pressure meter was used to assure that the speaker level was approximately the same to all the listeners (49-51 db). In addition, the amplifier (Gammal, model 82-1002) was equipped with filters and the signal from the tape recorder (Revox 716) was highpassed with the cutoff at 48 Hz.

An evaluation form was handed to each of the 4 listeners in listing of the evaluation form appears in Appendix II and separate instructions were given before each task. The instructions included playing the first 3 segments of each task as an example. This was done to avoid having to delete the first few segments of each task in the final tabulation of the results (the first few samples are generally eliminated because they usually do not reflect subject criteria). For the first task, the listeners were instructed to circle N on their evaluation form if the speech segment sounded natural or U if it sounded unnatural. The listeners were instructed, for the second task, to circle the closest in the pair (a,b) that sounded most like natural speech. For the third task, they were instructed to check off on their evaluation form the word same if the samples within the pair seemed to be identical, and the word different if they felt the samples were distinguishable.

Procedure for Pathological Cases

Vocal tract waveforms, derived from both normal and pathological speakers have been studied in a limited way (1,55-56). The limited amount of data collected is due primarily to the difficulties involved in recording the vocal cords. For the case of high speed photography the subject must tolerate a laryngeal mirror. The task of photographing a pathological case is made even more difficult due to the pathology itself. Usually, hours of

brackets are needed without any guarantee that the subject will expose his vocal cords. It is not easy to have to photograph a subject several times before obtaining an acceptable film.

In this research, one pathological case was filmed. The patient had cancer and as a consequence his epiglottis and ventricular folds (false vocal folds) were surgically removed. Following surgery, one of his vocal cords became totally paralyzed. Fortunately, the patient was able to tolerate extremely well the laryngeal surgery.

Using the patient's glottal area waveform, as measured from the high-speed film, synthetic speech was obtained using the procedure previously described. Problems were encountered in modeling his vocal tract due to difficulties in locating his formants. The results are presented in the next chapter.

Summary

The techniques that were developed in this research for the synthesis problem have been described in this chapter.

These are summarized in outline form as follows:

- i) The extraction of jitter and shimmer from the original speech waveform.
- ii) Filter design and modifications such as,
 - aa) Formant extraction and
 - bb) Formant bandwidth.
- iii) The modification of the radiation factor ($1-e^{-\alpha}$)

Most of the experimental work in this investigation was on a trial and error basis. Several methods for incorporating jitter and shimmer into the synthesized speech were tried, formant bandwidths were altered, various techniques to interpolate the glottal area waveform were tried, and the constant in the radiation factor $(1-\alpha^{1/3})$ was optimized based on a subjective criterion. It was not feasible to exhaust all possibilities. However, the procedure followed was not a random one and, whenever possible, theoretical explanations were always sought to describe experimental phenomena.

The method used to measure the glottal area was described in Chapter II. Basically, the system employed was unassisted whereby the operator traced the glottal opening using a specialized stylus, and the computer automatically calculated the area and stored it on disk. Other methods have been employed here (at the University of Florida) [34-36] and elsewhere [37-41]. Presently, work is being done here to fully automate the measurement of the glottal area from high-speed films.

CHAPTER 8

RESULTS

This chapter presents the results for both the normal and pathological cases. The first section illustrates the results of the synthetic speech for one of the normal subjects (S1). The reason why this particular case is presented here is because,

- i) the subject's fundamental frequency (124 Hz) was the lowest of all the 3 cases considered,
- ii) the subject was able to tolerate extremely well the largest mirror and as a consequence the film was not compressively dist. ed
- iii) the subject exposed a good portion of his vocal cords, thus maximizing the data for extrapolation in the measurement of the glottal area.

The second section contains the tabulation of the results from the listener evaluation of the synthetic speech. The last section presents the data obtained from synthesizing speech using the glottal area waveform derived from a pathological case.

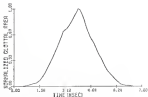
Synthetic Speech Results

We discussed in the previous chapter the fact that our glottal pitch period was a) periodically interpolated, b) periodically extended, and c) modified in interpolation (inter-

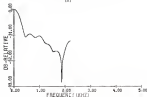
and glottal parameters. The residual waveform was used to synthesize speech.

The glottal area pitch period employed for subject JI and its corresponding FFT are shown, respectively, in Figures 3.1a and 3.1b. The area waveform shown in this figure was interpolated, thus, every frequency component after 3.2 kHz, for a source speed of 4,400 frames/s, is the result of the interpolation process. The interpolated glottal period and its FFT are shown, respectively, in Figures 3.1c and 3.1d.

The linear prediction filter was then designed from a 24 ms segment of the original speech record. From this model, the first 4 forward analysis poles of poles were extracted and a new filter formed. This was accomplished by factoring the linear prediction filter model and keeping only the forward residual poles. The frequency response of the linear prediction filter and the modified vocal tract filter are illustrated in Figure 3.2. The amplitude response of the linear prediction filter falls off at about 4 dB/octave faster than that of the modified model. The reason is that the linear prediction filter contains the overall effects of the glottal source and of radiation. The effect of glottal excitation is to produce a falloff of 12 dB/octave [11]. Radiation produces a rise of 4 dB/octave. Thus, the combined effects of both source and radiation, not present in the modified filter model, account for the additional fall in the frequency response of the predictor filter.

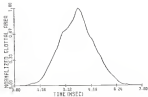


(a)

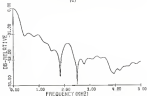


(b)

Figure 3-1 (a) Envelope of signal area waveform of reference cell. (b) FFT of above. A rectangular window was used.

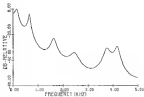


(a)

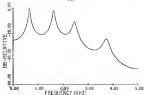


(b)

Figure 5.1 (a) Peak period of unfiltered electrocardiogram of subject 26; (b) FFT of above. A rectangular window was used.



(a)



(b)

Figure 3.2 Frequency response of a) 10th-order prediction filter of order $p = 10$, and b) modified vocal tract filter of order $p = 5$.

The method of linear prediction did not seem to give correct values for the formant bandwidths; thus, the bandwidths were changed. This is evident from a comparison of Figures 5.3a and 5.3b. The values used were taken from measurements of the vocal tract's parameters by Fujimura and Lindqvist [3]. Table 5.1 lists the first 4 formant bandwidths for both the linear prediction and the modified filter models.

It is also interesting to notice the first resonant peak, at about 150 Hz, in Figure 5.4c. This was due to a high energy content in the vicinity of the fundamental frequency in the amplitude spectrum of the real speech (see Figure 5.4a). Some linear prediction models use the envelope of the amplitude spectrum of the speech signal. It was not surprising to find this additional resonant peak.

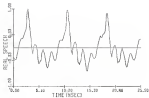
Table 5.1 Formant locations and bandwidths as derived from the linear prediction and modified filter models for subject JM.

	B_1	B_2	B_3	B_4
Formant Locations (Hz)	44	143	246	374
Linear Prediction Bandwidths (Hz)	43	154	205	184
Modified Filter Bandwidths (Hz)	48	70	135	184

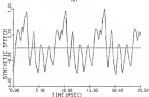
It was determined, by informal listening tests, that the value of α in the radiation factor $(1-\alpha r^2)$ that reproduced the most natural sounding speech for subject JM was at $\alpha = .7$. The value $\alpha = 1$, which is normally employed to simulate the effect of radiation, would not be used because this would have overemphasized the higher frequencies. The high-frequency components (greater than 2.2 kHz) were a direct result of interpolating the glottal area waveform. These components made the synthetic speech sound highly unnatural. A trade-off was then made to use the value $\alpha = .7$. For subject JM, at the cost of not fully interpolating the effects of radiation into the speech signal,

The natural speech and synthesized speech of subject JM are shown, respectively, in Figures 3.1a and 3.2a. Rather than comparing these two waveforms in the time domain, it is more enlightening to compare their spectra. The reason for this is that the natural speech had a significant amount of phase distortion as a result of the process of recording it, and playing it back in order to digitize it. Figures 3.1b and 3.2b show, respectively, the spectra of the real and synthetic speech signals. The following observations can be made:

- i) The 2 spectra, that is, the natural speech spectrum and the synthetic speech spectrum were similar up to 2.2 kHz. For frequencies beyond 2.2 kHz, they differed markedly. To approximate this better, the 2 spectra have been superimposed in Figure 3.3.
- ii) For frequencies greater than 2.2 kHz, the synthetic speech spectrum had less energy content than that of the real speech. This can be partially explained by the fact that the effect of radiation was not completely modeled.

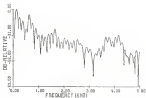


(a)

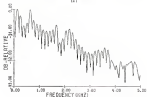


(b)

Figure 5.3 (a) recorded speech segment of subject JB
 (b) Synthetic speech segment of subject JB.



(a)



(b)

Figure 3-4 Spectrum of (a) natural speech, and (b) synthetic speech of subject 26. A Hanning window was used.

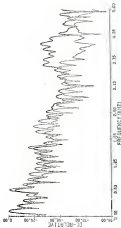


Figure 9.3 Spectrum of vocal (black line) and articulatory (light line) speech waveforms of subject CM. A windowing window was used.

- (ii) It can be hypothesized that the interpolation process was partially, if not completely, responsible for the above observations.
- (iii) A comparison of Figures 3.3b and 3.4b shows that the linear prediction filter did indeed model the envelope of the speech spectrum.

Appendix I gives, for subjects JB and JB, plots and tables of their data analogous to Figures 3.3-3.5 and Table 3.1, respectively. The observations and analysis of the data for subjects JB and JB are similar to the ones presented here for subject JB.

Listener Evaluation Results

The results for the first task are tabulated in Table 3.2. This task consisted of 40 segments of both the natural and synthetic speech from the 3 normal subjects. The listeners, upon hearing each segment, replied whether it sounded natural or unnatural. The values in Table 3.2 are the percentages of the times that the listeners judged the real and synthetic speech as sounding natural. The following observations can be made from the results of Task 1:

- (i) The real speech was not judged sounding natural in all the cases. In fact, for one of the subjects, JB, the real speech was judged as sounding natural only 4% of the time. The observed condition under which the subjects had no phoneme is a possible answer to explain these results. The subjects were required to phonate with a laryngeal mirror inserted in their mouths and to hold their tongues in a downward position.
- (ii) The synthetic speech was generally judged as sounding unnatural. The synthetic speech segment judged as the most natural sounding was that of subject JB (13.34 of the time).

Table 5.2 Task 2 results. Each value indicates the percentage of the times that the real and synthetic speech segments were judged as sounding natural.

SPEECH/SUBJECT	JR	MB	JA
Real	44.4%	41.7%	41.1%
Synthetic	15.7%	15.7%	11.1%

- (ii) It is interesting to observe that the order of preference for the speech segments as sounding natural was the same for both the real and synthetic segments. The order of preference was JR-MB-JA. Alternatively, as the quality of the natural speech increased, the quality of its synthetic counterpart also increased.

The results for the second and third tasks are both tabulated in Table 5.3. Both tasks consisted of playing pairs of speech segments. Each pair contained a synthetic speech segment and its corresponding natural counterpart. Task 2 was made up of 16 pairs of speech segments. The listeners, for this task, had to select which segment in each pair sounded more like natural speech. Task 3 contained 16 speech pairs, and the listeners' task was to reply whether the segments within each pair sounded the same or different. The results for these 2 tasks reveal the following (see Table 5.3):

- (i) In general, the listeners did not select the synthetic segment as sounding more natural than the real speech segment.
- (ii) The synthetic segment that scored the highest (55.1%) as sounding more natural than the real speech was for subject JR in Task 2. However, from the results of the first task, the real speech segment that sounded most natural was also that of subject JR.

Table 3.3 Results of Tasks 2 and 3. The values tabulated for Task 2 are the percentages of the times that the synthetic speech segment was selected from the speech pairs as sounding more like natural speech. Task 3 results are the per cent ratings of the times that the listeners judged the segments within the speech pairs to be different.

PAIR/SPEECH PAIR	20	80	10
iii Synthetic Segment	8.75	4.25	22.75
iii Different	87.25	1204	1095

- iii) Similarly, the synthetic speech and the corresponding natural speech segment, from the results of Task 3, did not sound the same.

The overall results from the listener evaluation tasks are tabulated in Table 3.4. In general, the results reveal that the synthetic speech did not sound natural and was distinct from the real speech. However, the real speech, in Task 1, was judged as sounding natural 71.54 of the time. The cause can be attributed to the abnormal conditions under which the sound was recorded during the filming session.

Table 3.4 Listener evaluation overall results.

TASK 1	TASK 2	TASK 3
Real speech judged as sounding natural: 71.54	Synthetic segment selected as sounding more natural than the real speech: 12.91	The segments within each pair sounded different: 89.14
Synthetic speech judged as sounding natural: 104		

Psychological Case Analysis

The vocal cord vibratory pattern of the pathological case, as measured from high-speed films, is illustrated in Figure 3.4. Various parameters, related to the area waveform, are tabulated in Table 3.3.

The shape of the glottal area waveform is seen to vary considerably from cycle to cycle. This could have been caused by the pathology itself or by operator approximation in actuating the glottal opening. The area waveform was characterized by a long closed phase ($OC = .84$). The subject's vibratory pattern, as observed from the high speed film, is described as follows:

The glottis opens simultaneously, along the full length of the cords. The right cord appears to travel further laterally than the left cord. The anterior section of the π [right] cord begins its medial movement before the posterior section or before the π [left] cord. Contact is made at the anterior and posterior sections, gliding back to the center. The cords stay closed for a very long period of time, approximately 23 frames. (18)

The speech waveform was also characterized by considerable changes in its detail from cycle to cycle (see Figure 3.4a).

Table 3.3 Glottal area waveform parameters, for the pathological case, averaged over 4 pitch periods.
 F_0 = fundamental frequency, OQ = open quotient,
 OP = open period, CP = closed phase and CF = closure period.

F_0 (Hz)	OQ	OP (ms)	CP (ms)	CF (frames/s)
128	.85	4.8	5.1	4108

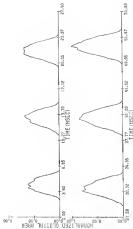


Figure 5.4 Head and elementary patterns at pathological subjects as measured from high-speed films.

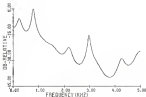
The reason for this is that it was difficult for the subject to sustain phonation over a long period of time at a constant intensity and frequency. Thus, there was no guarantee that the film segment picked for analysis correlated with the speech segment recorded previous to the filming.

A 10 ms segment of the original speech was picked for analysis in order to determine the vocal tract filter coefficients. Figure 3.1a shows the frequency response of the linear prediction filter of order $p = 10$. The problem was to determine the formants. Table 3.4 lists the values employed and the frequency response of the modified vocal tract filter is shown in Figure 3.1b.

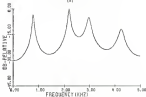
Table 3.4 Formant locations and bandwidths as derived from the linear prediction and modified filter models for the pathological subject.

	F_1	F_2	F_3	F_4
Formant Location (kHz)	.784	3.13	3.98	4.28
Linear Prediction Bandwidth (Hz)	90	120	75	228
Modified Filter Bandwidth (Hz)	10	80	110	175

The speech was synthesized in the same manner as before. That is, one glottal pitch period was selected (the fifth cycle in Figure 3.8i), it was periodically extended, the pitch and glottal parameters, obtained from the original



(a)



(b)

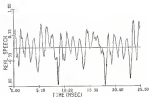
Figure 8.7 Frequency response of a) linear prediction filter of order $p=18$, and b) modified vocal tract filter of order $p=8$ of pathological subject.

speech waveform, were incorporated into the glottal area waveform; this waveform was used to modify the vocal tract filter, whose frequency response is shown in Figure 3.1b, and finally, the effect of radiation was incorporated by filtering the output of the vocal tract model with the filter $(1-.3e^{-1/2})$.

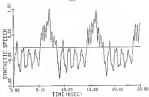
The synthetic speech obtained as a result of these operations is shown in Figure 3.1b. Figure 3.1a shows a segment of the original speech. Again, as in the previous case, the comparison between the synthetic speech and the natural speech is made in the frequency domain. The following observations can be made based on the results presented here and on informal listening tests:

- i) The spectra of both the synthetic and real speech, shown in Figure 3.2, differed substantially. As mentioned previously, they have been superimposed in Figure 3.1c.
- ii) The above differences can be attributed to:
 - a) The operation used to extrapolate the pre-burster portion of the glottal area opening in order to measure the area.
 - b) The problem of instant frequency estimates from the linear prediction model (refer to Figure 3.7).
 - c) The real speech and the file were probably not well correlated.
- iii) The subject's speech was characterized by large variations in closure and jitter from cycle to cycle. When these parameters were incorporated into the synthetic speech, it gave it the same harsh quality found present in the real speech.

The next chapter discusses the implications of these results.

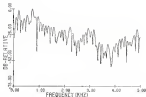


(a)

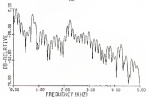


(b)

Figure 3-8 (a) Natural speech segment of pathological subject;
 (b) Synthetic speech segment of pathological subject.



(a)



(b)

Figure 3-7 Spectrum of a) natural speech, and b) synthetic speech of pathological subjects. A Hamming window was used.



Figure 3.23 Spectrum of real (black line) and synthetic (light blue) speech of pathological subject. A smoothing window was employed.

CHAPTER VI

DISCUSSION

The steps to the objectives of this research are presented in Figure 1.3, which is shown again here (Figure 1.1) for convenience.

The global area waveform, extracted from high-pass slices of the vocal sounds, was used to create a digital filter model of the vocal tract. The synthesized speech derived from the speech production model was then compared to the original speech recorded at the filming session. A formal listener session was held to judge the synthetic speech in terms of its naturalness and to compare the synthetic speech with the real speech obtained at the filming session.

The results from the formal listener session showed that the synthetic speech was not perceived as natural. When presented in pairs (synthetic speech versus its natural counterpart), the listeners preferred the natural over the synthetic speech 84% of the time. When the segments were presented individually, the listeners judged the synthetic speech as sounding natural 18% of the time, and the real speech was judged natural 81.3% of the time.



Figure 4.1. Flow diagram illustrating the steps leading to the research objectives.

Based on the results of the listener evaluation, the following conclusions were reached:-

- i) The synthesized speech and its natural counterpart were indistinguishable.
- ii) The synthetic speech derived from the speech production model employed in this research did not sound natural.
- iii) The real speech also suffered from a lack of naturalness. This could be attributed to the abnormal conditions under which the speech segment was recorded.

The implications of these results with regard to the glottal area, synthetic and real speech, and the pathological case are discussed in due chapter. The last section discusses

the conclusions reached on the adequacy of synthetic speech derived from glottal area excitation and given recommendations for future research.

Glottal Area Waveforms

The results of the synthetic speech derived from glottal area waveform excitation leads one to conclude that the area waveforms may not contain all of the parameters necessary for the production of natural speech. Alternatively, the differences between the area waveforms and the corresponding volume velocity may be significant in terms of characterizing the detail which contributes to the naturalness of spoken speech.

It can be argued that the glottal area measured from high-speed films did not reflect the true glottal area waveform because:

- a) The anterior-posterior portions of the vocal cord aperture were often misjudged by the speaker.
- ii) Operator approximations in the measurement of the glottal area were made due to scaling film, notes on the vocal cords, and uncertainty as to the location of the glottal edge.
- iii) The camera sampling rate of approximately 7100 frames/s was inadequate. Thus, the high frequency components, which contribute to the detail of the area waveforms, were not present.
- iv) A short segment was not an adequate representation of the glottal area waveform. The process of periodically extending a glottal cycle, and hence perceiving the jitter and duration parameters derived from the original speech could have deteriorated the synthetic speech. It would have been more realistic to measure one second of glottal activity and use this to create a vocal tract model, but this was not possible.

- vi) Interpolating the data was found introduced higher frequency components which made the speech sound unnatural. To investigate the effects of the interpolation process, one pitch period of volume velocity was obtained by inverse filtering the sampled ($f_s = 48$ kHz) speech waveform. This volume velocity cycle was then resampled at 8.1 kHz and periodically interpolated to increase the sampling rate to an effective 10 kHz. Both volume velocity cycles, that is, the original and the processed, were then used to synthesize speech employing the same technique as for the glottal area cycle. When compared to the original speech, the synthetic speech derived from the processed volume velocity cycle had lost some of its natural qualities than the speech derived from the unaltered volume velocity cycle.

The above arguments are all based upon the fact that some temporal detail in the area waveform was lost or altered by processing the data. However, current research in vocal cord modeling (refer to Appendix B) indicates that the glottal area waveform is not characterized by temporal detail. From their two-mass model of the vocal cords, Ishizaka and Flanagan [17] obtained glottal area and volume velocity (glottal flow) waveforms which were distinct in their detailed structure. The differences observed can best be described as follows:

The glottal flow wave is typically characterized by some temporal detail, asymmetry, and steep falling slope, while the area wave shows little temporal detail, is less steep, and tends to be more symmetrical because the cords are moving and are generally forced at a frequency above their frequency; their mechanical displacement does not reflect the detail of acoustic interaction which the glottal flow displays [17, pp. 1894-1898].

Thus, the differences between the volume velocity and area waveforms derived from the two-mass model can be attributed to source-filter interaction.

The uniqueness of their model was that it incorporated aeromechanical interaction and it was based on 8 physiological constraints, additional pressure and tension; rest area of glottal opening; and vocal tract shape. These parameters were derived from experimental measurements done on the human vocal production. Then, these data had to be incorporated into a model all the parameters necessary for the production of natural speech, which included source-tract interaction.

Other related research has been done in vocal cord modeling and similar observations of the differences between the area and volume velocity waveforms have been made [43-44]. From the results of vocal cord modeling, the effect of tract excitation is evident on the volume velocity and not on the area waveform. Some researchers argue [1,45] that it is important, if one is interested in the production of natural speech, that the synthesis system includes source-tract interaction.

If it is true that the vocal cord vibratory pattern is insensitive to tract interaction, then it can be concluded that the area waveform is not adequate for a complete description of voiced speech.

Synthetic And Real Speech Research

There were several factors that played a significant role in the naturalness of the synthetic and real speech. These factors consisted of the instrumental efforts on the

naturalness of the speech signal due to the processing of the data, and to the vocal tract model employed in that research.

The listener position model was designed by analyzing a sampled version of the original speech recorded during the filtering session. The formant frequencies were estimated, the formant bandwidths altered, and the nonlinear factor ($1-\alpha^{-2}$), $\alpha_{\text{avg}}=75$, implemented. Synthetic speech was obtained by driving this speech production model with the glottal area waveform. A tape, consisting of both synthetic and real speech, was prepared for the listener evaluation tests.

The effects of these steps on the speech are summarized as follows:

- K) Both the natural and synthetic speech had a low-pass filter. This was obtained by low-pass filtering the natural speech at 4 kHz before sampling. For the listener evaluation tape, the speech (natural and synthetic) was low-pass filtered at 4 kHz before being recorded on tape.
- L) The listeners rated the low-pass filtered synthetic speech. This could be another possible explanation as to why the real speech did not sound highly natural since the higher formants were not present.
- M) In most speech synthesis models, as is the one employed in this research, only the first 3 or 4 formants are employed. These first formants account for the perceptual intelligibility of speech, whereas the higher order formants contribute to its naturalness [1]. The effect of the higher order formant poles was not present in the synthetic nor in the altered speech.
- N) There is uncertainty as to the effect of different formant bandwidths. The literature to date does not seem to discuss any research in this area. The formant bandwidths, as derived by the method

of linear prediction, did not contend with the present literature with assumptions made on human. The conclusions were then altered to conform to experimental measurements by Fujisaki and Lindqvist [10].

- v) The process of interpolating the glottal area as an alternative to the had a deteriorating effect on the naturalness of the synthetic speech. This was discussed in the previous section.
- vi) The effect of radiation can be simulated by the term $(1-s^{-2})$. This is an approximation valid for frequencies less than 4 kHz. The effect is to intensify the higher frequency components. The higher frequency components in the synthetic speech were a result of the interpolation process. Applying the term $(1-s^{-2})$ served only to brighten these components, giving the synthetic speech a more artificial quality. A tradeoff was made to use $(1-s^{-2})$, viz. 75. Thus, the synthetic speech did not obtain the full effect of radiation.
- vii) In natural speech, the excitation (represented) is characterized by the speech wave gradually increasing (decreasing) in amplitude. For both the synthetic and natural speech sequences, there was simulated by multiplying the onset and offset by a linear 20 ms ramp. This would have contributed to the naturalness of both the real and synthetic speech signals.
- viii) The speech production model employed in this research was linear and did not incorporate source-trait interaction. It is still unclear as to the nature of source-trait interaction on the naturalness of speech. However, from the current research on vocal cord modeling (prior to previous version and Appendix A), source-trait interaction seems to be an important factor contributing to speech naturalness.
- ix) In general, the vocal tract consists of both poles (resonances) and zeros (antiresonances). However, for voiced unnasalized vowels, it is well known that the vocal tract exhibits only poles [2]. Thus, the fact of approximating the vocal tract by an all-pole filter model seemed to be adequate for this research.
- x) The additive parameters of jitter and shimmer are continuous functions randomly distributed in the speech signal, that is, the jitter and shimmer do not just occur at one time instant in the glottal

were, but rather are distributed throughout the entire word. It can be argued that the manner in which they were extracted from the original speech and incorporated into the synthetic speech (order vs. duration) did not fully characterize these parameters. However, little work has been done in studying these parameters, and it is still uncertain as to which method of analysis works best in extracting the jitter and duration parameters.

Undoubtedly, the synthetic speech suffered from a lack of naturalness due to the above stated reasons. Moreover, it seems that the above factors which contributed most to the unnaturalness of the synthetic speech were: the localization of the glottal area (which was required due to low camera sampling speed); the absence of nonverbalized intonations; and the lack of the perceptual effect of higher order formant poles.

Although the natural speech was processed, the factor that probably influenced most its naturalness was the insertion of a laryngeal mirror into the subject's mouth at the time of filming. At the present, there seems to be no way to overcome this obstacle, except by employing other techniques to measure the glottal area.

Psychological Case Studies

Originally, the vocal cords of two pathological cases were filmed using high-speed photography. Both subjects were able to tolerate well the laryngeal mirror. One of the subjects (CC) had cancer and as a consequence his epiglottis and ventricular folds were surgically removed. Following

surgery, one of her vocal cords became paralyzed due to squamous metaplasia.

The second subject (SH) had an acute upper respiratory infection. The data, however, was not clear. This made it difficult for the operator to measure the glottal area. Furthermore, the vibratory pattern of the vocal cords was out of phase, that is, as one cord was making a lateral excursion, the other was making a medial excursion. Thus, the operator was never certain when the maximal glottal opening occurred. For these reasons, this data was not considered quantitatively in this study.

The treatment of the pathological subject's data followed closely that of the data derived from the normal subjects. The synthetic speech was obtained by considering one pitch period of the glottal area waveform (fifth cycle in Figure 5-6). The pitch period was periodically extended and the graver and thinner formants incorporated. The formant bandwidths were altered in a manner similar to that used for the normal subjects. For this study, the experimental data by Fujisaka and Luskquist (30) as formant frequency and bandwidth were used for the normal subjects. However, no related information on pathological subjects is known to exist. This complicated the analysis and introduced much trial and error in attempting to arrive at similar formant bandwidths.

The synthetic speech derived by filtering the glottal area waveform with the vocal tract model is shown in Figure 5-8a. The original speech is shown in Figure 5-8b.

The subject [EC] could not sustain phonation at a constant frequency and intensity level, and as a consequence the speech was highly irregular from cycle to cycle. Thus, there was no assurance that the film was correlated with the sound recording. In fact, the analysis of the film showed that the visual words stopped vibrating after approximately 15 feet of film.

As a result of his inability to sustain a constant phonation, difficulties arose in extracting the formant frequencies. Figure 4.3 illustrates the frequency response of linear prediction filter (p-18) for two different 15 ms speech records of the pathological subject. The records were separated by 15 ms. A comparison of the two separate period graphs shows that they differ in formant structure, that is, in formant bandwidth, content and location. Hence, the task of formant extraction became one of guesswork.

A comparison between the glottal area waveform (Figure 4.4) and the original speech (Figure 4.5a) results in the following observations:

- i) Both the glottal area and speech waveforms were characterized by large variations in shape from cycle to cycle.
- ii) Both waveforms were characterized by a significant amount of jitter and shudder.

The synthetic speech derived from the glottal area waveform differed in sound from the original speech. However, both were characterized by the same harsh sound which is known to be due to jitter and shudder [10-11]. From this it seems that the method employed to extract these parameters from the speech waveform was adequate.

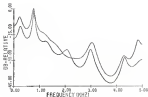


Figure 4-8. Frequency response of linear prediction model. (p-18) for two different 20 ms speech segments of the pathological subject.

Finally, we remark that much work remains to be done along the lines of Ferguson and Lisowski [3] for pathological cases if one is to be successful in synthesizing pathological speech which sounds undistorted and thus similar to the original phonation.

concluding remarks

This study was based on investigating whether the glottal area waveform was sufficient for the production of natural speech. From the results of the listener evaluations we can conclude that this is not the case.

The question that arises is if it is worthwhile to synthesize speech using the glottal area waveform as an excitation function in a speech production model. Alternatively, can more insight be gained from the area waveform or from the derived synthetic speech? It seems that the glottal area is a better vehicle for studying the vibratory process than the derived synthetic speech. The reason is that the simple nature of the area waveform, as compared to the speech wave, allows a quicker assessment in trying to determine abnormal vocal vibratory processes. Additionally, to synthesize speech the vocal tract formant structure must be derived. In the case of a pathological subject, this can be an almost impossible task.

The method of high-speed photography, although more costly and time consuming than other methods to measure glottal activity, such as optical glottography, electrical glottog-

graphy, stroboscopy and inverse filtering, has various advantages (Saler in [7] has a brief summary on the principal techniques for investigation of laryngeal activity). These advantages are outlined as follows:

- (i) The vibratory pattern of each individual vocal cord can be extended. Other parameters, such as glottal width, length, and motion, can also be determined.
- (ii) The advantage of high-speed photography, not common with other methods, is the opportunity to assess the vocal cords directly. Abnormal conditions, such as lack of phase vibration of the vocal cords, can easily be determined. From the glottal area waveform itself, it is not possible to determine with certainty, asymmetrical vibrations of the vocal cords. This was the case for one of the pathological subjects (SM).
- (iii) Black and white film was used for this study to photograph the vocal cords. However, color high-speed film is also available, allowing a better enhancement of the detail of the cord structure. This is especially an aid in studying and diagnosing pathologies.
- (iv) Other phenomena, such as cough and laughter, can be studied by this method.

The glottal area does not seem to contain all the information relevant to certain voice disorders (e. g. the examples, asymmetrical vibrations of the vocal cords or perturbed vocal cords). Thus, we can conclude that the glottal area waveform can be used as a parameter to determine a pathology if used in conjunction with the visual assessment of the high-speed film. This assessment is not possible with the other methods available to measure glottal activity.

For future research, it is suggested that the following areas be pursued further:

- i) The development of faster and more efficient ways to measure the glottal area from high-speed films.
- ii) The synchronization of the film with the sound recording. Presently, this is not possible due to the high camera noise.
- iii) Photograph the vocal cords at the rate of 14,000 frames/s. This will avoid the need to interpolate the glottal area if one is interested in synchronizing speech. The Fastax camera employed in this research can obtain a speed as high as 8,000 frames/s. However, to photograph at this speed special lighting and film, not available for this research, are required.
- iv) A comparison of the glottal area as derived by other methods. One such study was made comparing the glottal area measured from high-speed films to that obtained from photoelectric spectrography [4].

APPENDIX B

SPEECH SYNTHESIS BASED ON VOCAL TRACT MODELING

The purpose of this section is to review vocal tract modeling as a tool for speech synthesis. For a tutorial review on vocal tract modeling, the reader is referred to NAR (1981).

Most present-day techniques make the assumption that the source and vocal tract are linearly separable. This simplifies analysis and at the same time leads to workable models requiring few parameters to completely describe them. In fact, many schemes, such as linear prediction (previously discussed), include the effects of glottal excitation into the vocal tract filter. This makes it possible to produce voiced speech by driving the filter with a series of impulses separated by the duration of the pitch period. Although the speech obtained from these methods, which assume linear separability between source and tract, is highly intelligible, it sometimes lacks the natural quality found in real speech. In order to overcome this the design should include the details of our vocal mechanism that are responsible for natural speech production. This means that all of the parameters which play a significant role in the production of human speech should be modeled.

Attempts to model the vocal cords, for the purpose of synthetic speech production, date as far back as the late 1880's when Alexander Graham Bell with the assistance of his brother Melville constructed a wax model of the human head. Their model included lips, tongue, palate, teeth and alveolars. The vocal cord orifice was made by stretching a stretched rubber sheet over tin supports. They were able to produce simple utterances, and although they probably were not aware, their model incorporated the effects of source-tract interaction.

The first quantitative self-exciting one-mass model of the vocal cords was by Flanagan and Landgraf [14] in 1948. The model came about as a direct result of previous work, by Flanagan himself, in attempting to model the vocal cords [15-17]. It was implemented to produce synthetic voiced speech [18] and later his vocal cord/vocal tract speech synthesizer was improved to simulate voiceless excitation [14]. Another investigation, Crystal, also produced a similar model [44]. However, his model did not account for the closed phase of the glottal pulse period. He thus limited himself to analyzing the open period of the glottal cycle. As a consequence, his model was inadequate for speech synthesis.

Flanagan's one-mass oscillator model of the vocal cords is shown in Figure 8-1. In this figure P_0 is the subglottal pressure, P_1 and P_2 are, respectively, the inlet and outlet pressures, d is vocal cord thickness, l is cord length and A_0 is glottal area. Two masses are present, one for each vocal



Figure 4-1. Second-order mechanical model of the vocal cords (after [141]).

good. Since bilateral symmetry is assumed, calculations need be made only for one vocal fold (mass). Motion of the mechanical oscillator is restricted to one dimension and can be described by the second-order linear differential equation

$$M\ddot{x} + R\dot{x} + Kx = F \quad (4-1)$$

where $x = x(t)$ is the mass displacement, $F \equiv F(t)$ is the system's forcing function, M is the oscillator mass, K is the spring stiffness constant and R is the viscous damping of the system.

The driving function $F(t)$ is determined from pressure variations along the model. If we first consider the idealized case, when all losses can be neglected, we have from Figure 4-2a

$$P_1 = P_2 = -P_B \quad (4-2)$$

P_B is the Bernoulli pressure given by



Figure A-2 Distribution of pressure within the glottal orifice: (a) ideal, (b) actual (after [18]).

$$p_g = \frac{\rho}{2} \pm \frac{v_g^2}{2} \frac{A_g}{A_g} \quad (A-3)$$

where $v_g = v_g(x)$ is the volume velocity through the glottal orifice, $A_g = A_g(x)$ is the glottal area function and ρ is the density of the air. In the actual case the abrupt transition from the large area of the trachea into the relatively small glottal inlet causes a contraction of the entering jet known as a vena contracta. The effect is that the effective area of the orifice appears to be smaller than it actually is, consequently the actual pressure drop at the inlet to the glottis will be greater than that discussed under ideal conditions. Furthermore, loss effects are also present when the area increases suddenly at the outlet of the glottal orifice. Experimental measurements by van den Burg et al. [18] led to the results shown in Figure A-2b. We see that

$$P_1 = P_2 = 1.27 P_0] \quad (2-4)$$

and

$$P_2 = -0.5 P_0 \quad (2-5)$$

where P_0 is given by (2-1).

Flanagan, for his one-mass model, took F to be equal to the mean inlet and outlet pressures acting on the cross-sectional surface area. In equation form we have

$$F(t) = \frac{1}{2} (P_1 + P_2) A(t) \quad (2-6)$$

Initially the cords are at rest ($x=0$) with the glottal surface having minimum area. Defining this as the rest area A_{g0} gives

$$A_g(t) = A_{g0} + ix \quad (2-7)$$

When motion of the cords is initiated, that is, when F is applied, a point will be reached where the two masses (cords) will touch and the glottal area (and hence volume velocity) will be zero. The value of x for which this occurs is $x_0 = -A_{g0}/i$, where x_0 is referred to as the critical value of x . Two different situations are then analyzed, the open period and the closed phase. Taken together they constitute the total glottal cycle.

The open period is characterized by sinusoidal oscillations justifying the spring constant to be chosen so as to produce an undamped natural frequency F_0 of the vibrating system well below the fundamental frequency when forced. For

has steady vibrations whose $\dot{x}_0 = 13$ in. $\ddot{x} = .24$ gram, $\ddot{x} = 4 \cdot 10^3 \text{ cm}^2/\text{sec}^2$ and $\ddot{x} = 5$.

The closed phase represents a totally different situation. When the two masses collide two distinct cases can be hypothesized. The masses can be infinitely hard or purely viscous. In the first case, upon collision, the forcing function takes the value $\frac{1}{2} F_0 \delta(t)$. This value is instantaneous since upon contact the forces acting on the masses will be in a direction to open the orifice. The closed phase will therefore be of short duration and for this case we use that $\ddot{x} \leq \ddot{x}_0$. The second case represents the more realistic case where the vessel walls actually squeeze into one another owing to their compliant factor. The forcing function at closure is $\frac{1}{2} F_0 \delta(t)$ and will be maintained until $\ddot{x} = \ddot{x}_0$ at which time the forcing function will again change to some other value. The closed phase will then be determined by the total time at which $\ddot{x} \leq \ddot{x}_0$. During closure the damping constant of the system will be different from the value at the open period. Its value will be equal to the sum $\ddot{x} + \ddot{x}'$ where $\ddot{x}$ is the damping constant during the open period and $\ddot{x}'$ is the damping constant of the opposing mass. Plesinger assumed $\ddot{x}' = 1/\sqrt{2}$, the condition for critical damping.

Since in real space $\ddot{x}$, $\ddot{x}$ and $\ddot{x}$ are continuously changing, Plesinger defined a good measure parameter Q by the relationship

$$Q^2 = \left(\frac{1}{\ddot{x}}\right)^2 \frac{\ddot{x} \ddot{x}'}{\ddot{x} + \ddot{x}'} \quad (2-4)$$

in order to combine the stiffening and lightening of the dynamic system. The effect of q is to scale the system parameters as seen in (A-6):

Now that the vocal tract model has been described as proposed in diagram its operation as the glottal excitation to a vocal tract model, Figure A-3 shows the network representation of the vocal system for voiced sounds as proposed by Flanagan. The three parameters $R_q = R_q(0)$, $R_k = R_k(0)$ and $L_q = L_q(0)$ represent respectively the viscous nonlinear-dependent resistance, the kinetic flow-dependent resistance and the inductance owing to the mass of the glottal air plug. These terms, which were defined in accordance with van den Burg's previous work [18], are as follows:

$$R_q = 1.6 \times 10^{-3} / A_q^2 \quad (A-6)$$

$$R_k = .447 \times 10^{-3} / A_q^2 \quad (A-7)$$

$$L_q = .05 / A_q \quad (A-8)$$

where ρ is the kinematic viscosity of air and the other terms are as previously defined. The subglottal pressure is approximated by the variable battery P_q . This is justified by the fact that the lungs appear as a low impedance constant power wave source when viewed from the vocal cords [2].

The vocal tract was modeled by an articulatory synthesis algorithm. Each T-section in Figure A-3 represents a portion of the vocal tract having cross-sectional area $A_i = A_i(0)$. The

The results from Finnegan's nonlinear model can be summarized as follows [14-16]:

- i) The fundamental frequency of the vowel sounds was highly dependent upon vowel length configuration. Variations in the order of 10 Hz were sometimes observed between two different vowels, all other conditions being held constant.
- ii) An increase in air pressure produced a lower fundamental frequency.
- iii) The glottal area waveform was usually more smooth and symmetric than its corresponding volume velocity waveform. In fact, at times they differed substantially in form.
- iv) Source-tract interactions played a significant role in the shape of the volume velocity waveform and had little influence on the overall shape of the area function.
- v) Some temporal details were identified in the volume velocity waveform as being due to the first formant frequency of the vocal tract. This did not hold true for the area function, confirming the previous result.

The one-mass model, as described above, could not sustain oscillations whenever the vocal tract was constrained at some point (that is, for some i such that $A_i = 0$). A later modification of Finnegan's nonlinear model by Steinlechner [17] overcame this. The model is illustrated in Figure 4-4. The difference between this and the former one-mass model is the addition of a second spring K_2 of length x_2 when unstretched. The spring K_1 can be thought to simulate the unstriated and striated cricothyroid muscles and K_2 the vocalis. For small x the effect of spring K_2 can be superimposed on that of K_1 and the system becomes identical to our original one-mass model. Setting of the oscillator is again one dimensional and is described by the nonlinear system



Figure 3-8 Schematic diagram of two-spring-mass system modeling vocal cord movement (after [81])

$$m\ddot{x} + kx = k_1x + (k_2/c_0^2)x^3 = f(t)$$

where k_1 and k_2 are the spring constants of k_1 and k_2 , respectively, and $f(t) = [D_0(D_1/2 + D_2)]$. The other parameters have already been defined and are the same as in the one mass/one spring system.

The model was implemented using the model system shown in Figure 3-1. The results were identical to Flanagan's model except that oscillations were sustained during inter-vocalic voiced stop production.

The two-mass model of the vocal cords was able to simulate many important features of the human vocal cords. The three basic output parameters, glottal area function, volume velocity and the speech waveform closely resembled those associated with actual speech. However, the model was inade-

quite in its inability to produce other physiological detail. For example, it could not produce different register nor could it account for phase differences within the vocal cords. The model was also not able to sustain smaller tones under a resistive output load. As a consequence phonation was not possible for frequencies near the first formant since in this region the vocal tract's input impedance is negative [11]. It also suffered from too much source-tract interaction reflected by the high variations in fundamental frequency dictated by changes in vocal tract configuration.

The development of a two-mass vocal cord model came about as an effort to overcome these difficulties. An attempt by Hodgson [10] to simulate such a model proved its feasibility. However, the model had certain flaws due to design criterion. In 1948 Ishizaka and Matsudaira [19] designed a two-mass model of the vocal cords similar in principle to Flanagan's one-mass model and was later computer-implemented by Ishizaka and Flanagan [12]. The model has been used in a speech synthesis system by Flanagan *et al.*, [18] and more recently has been implemented so that its control parameters are derived from printed text [16].

Figure A-5 gives the schematic diagram of the two-mass approximation of the vocal cords. The masses are only allowed lateral displacement and their movement can be described by a pair of second-order nonlinear differential equations with time-varying coefficients. The spring k_p



Figure A-3 Schematic diagram of the two-mass approximation of the vocal cords after [17].

coupled the two masses together and is analogous to the vocalis muscle. As k_g tends to infinity the model becomes essentially a one-mass model. The springs x_1 and x_2 are analogous to the action of the anterior cricothyroid muscle and other muscles. As with the one-mass model, the two-mass model's behavior can be characterized by a set of four input parameters. These parameters are identical to the one-mass case, that is, subglottal pressure, cord location, rest area of cord opening and vocal tract shape. Electrical impedances were derived from the pressure drops along the orifice and a vocal system, similar to that shown in Figure A-3, was implemented for the synthesis of voiced sounds.

The major results, as obtained by Ishizaka and Flanagan [17], are outlined below:

- 1) The fundamental frequency was almost independent of the vocal tract shape in contrast to the one-mass model, which was not. Two reasons seem to account for this. First, was the more symmetrical nature of the two-mass model in

describing the human vocal cords. A second factor was that (a-10). That is, $P_{12} = -.5 P_0$ was not used to determine the pressure drop at the outlet of the glottal orifice. Instead, the value of P_{12} was calculated from fluid mechanics theory and not from van den Berg's results. The value obtained was $P_{12} = P_1 = -CP_0$, where $.0002748$, which was less than van den Berg's figure of $.5$. This led to a more realistic simulation of coupling between source and tract.

ii) The open quotient (open period of glottal waveform over its total cycle) decreased with an increase in sub-glottal pressure until reaching a value of $.8$, at which point it remained fairly constant, even with further increases in the subglottal pressure.

iii) By proper selection of the model's input parameters the open quotient was made to increase with the fundamental frequency.

iv) The differences between glottal area and the corresponding volume velocity wave can best be summarized by a direct quotation from Ishizaka and Flanagan's paper describing the model's implementation:

The glottal flow wave is typically characterized by more temporal detail, asymmetry, and steep falling slope, while the area wave shows little temporal detail, is less steep, and tends to be more symmetrical. Because the cords are passive and are generally forced at a frequency above their natural frequency, their mechanical displacement does not reflect the details of acoustic interaction which the glottal flow displays. [17, pp. 1294-1295]

v) Similar to the case-case case, volume velocity waveform was affected by changes in the vocal tract configuration, this not being the case for the area waveform.

vi) A change in fundamental frequency, for a fixed vocal tract configuration, had no significant influence on the shape of the glottal area waveform. However, marked changes were observed in the corresponding volume velocity waveform, especially for low fundamental values of the vocal tract.

vii) Increasing the cord tension parameter Q decreased the amplitude of glottal area and volume velocity while not affecting the open quotient, phase difference between the two signals, and the shape of the glottal area. The volume velocity waveform was however affected, since an increase in Q increased the fundamental frequency.

The overall advantage in vocal cord modeling is the attainment of source-tract coupling, not possible in linear systems which assume linear separability between source and tract. As a consequence a great amount of physiological detail, corresponding to the human vocal structure, is achieved. To investigate the effects of source-tract interaction Garcia et al. [42] recently proposed the model shown in Figure A-4. The two resonant circuits on the right hand side simulate the driving point impedances at the vocal tract. The different parameters are defined as follows: $F_1(\Omega)$ = first formant frequency, $F_2(\Omega)$ = second formant frequency, $L_1(\Omega) = \kappa_1/\pi F_1^2(\Omega)$, $\kappa_1(\Omega) = \kappa_1 F_1^2(\Omega)$ and $L_2(\Omega) = \kappa_2/\pi F_2^2(\Omega)$. ϕ_1 , ϕ_2 and β_2 are fixed, κ_1 , κ_2 and L_0 are defined by $G/H = G_1/H_1$, respectively, and $k(t)$ is obtained by either the one- or two-mass model of the vocal cords. We see that the two



Figure 3-4 Motor source model loaded by an equivalent circuit of the main input impedance (after [8]).

mainstream circuits are controlled by the first two resonant frequencies. Since the present study only investigated the effects of first-resonant frequency on volume velocity, the second resonant circuit can be neglected. The circuit in Figure 3-4 can be described by a system of nonlinear differential equations with constant coefficients. The solutions were obtained by the Runge-Kutta method ensuring more precision, as compared to having used difference equations. In Plesinger's model the term $d/dt[L_y f(t)]$ was approximated by $L_y df(t)/dt$. This study considered both cases in order to see the effect of this approximation. The results are outlined below.

- (i) When the approximation to $d/dt[L_y f(t)]$ was used the effects of brass loading on Helmholtz volume velocity were similar to that obtained by Jelinek and Plesinger.
- (ii) Using the exact value of $d/dt[L_y f(t)]$ accentuated even more the coupling effects of source-circuit on the volume velocity waveform.

- iii) The effect of source-trait interaction on the amplitude spectrum of volume velocity was one of adding zeros near the vicinity of the formant frequencies.
- iv) One of the major causes of source-trait interaction was the fact that the input impedance of the vocal tract attains its maxima at the formant frequencies and is equal to that of the glottal source in the vicinity of the first formant. Outside this range the vocal tract's input impedance is small and its effect on the source is comparatively negligible [1].

Further research is being conducted using this model and it has recently been hardware-implemented [4].

A two-dimensional model of the vocal tract was recently proposed by Tjvitz [70]. His model differed from Ishizaka and Flanagan's two-pass model in that he divided each vocal tract into two vertical levels, as compared to the formers' horizontal levels. Each vertical level consisted of eight equal masses. In Tjvitz's model movement of the masses was two dimensional, that is, the masses could move both horizontally and vertically. One level was analogous to the nasopharynx and the other to the oropharynx. His results seemed to indicate source-trait interaction in the volume velocity waveform but not as pronounced as in the two-pass model. No comparisons can be made between area function and volume velocity since his study did not discuss this. The speech obtained from the two-pass model, however, seems to be of much better quality than that from the six-pass mass model.

In the area of applications present efforts are being directed towards an automated speech synthesis system con-



Figure 8-7 Automatic generation of control functions for the non-linear synthesizer from printed text input (after [18]).

modeling of a two-man model whose input parameters are derived from printed text. Figure 8-7 shows the block diagram implementation for such a model. The model is capable of voice-excited speech synthesis with vocal coupling (represented by the parameter K in the figure). Although still in the experimental stage results indicate striking future perspectives for speech synthesis based on vowel sound modeling.

APPENDIX B

METHOD OF LINEAR PROJECTION IN SPEECH PRODUCTION

The recent works of Makhovskii, Reiz, Markel and Orlov (48,49,50,51,52) describe the method of linear projection. The following is a general outline of the theory.

In vocal tract modeling, the main objective is to obtain a transfer function having characteristics similar to the human vocal tract. Speech, because of its fluctuation in pitch and vocal tract shape, is a nonstationary phenomenon, that is, its characteristics vary with time. Nevertheless, if we consider a small segment of speech on the order of 10-20 ms, an interval over which the vocal tract does not change appreciably in shape, then we can model the human vocal tract by a linear time invariant filter. The diagram given in Figure B-1 illustrates a speech production model. If $x(t)$, the vocal tract transfer function is given by



Figure B-1. A speech production model.

$$W(z) = K \frac{1 + \sum_{k=1}^P b_k z^{-k}}{1 + \sum_{k=1}^Q a_k z^{-k}} \quad (8-1)$$

where $z = e^{j\omega T}$, $\omega = 2\pi f$ is a complex frequency variable, $T = 1/f_s$ is the sampling period, f_s is the sampling frequency and K is the gain. The input $g(nT)$ is the digital excitation and this need not be unique. However, the model, as derived from the linear prediction method, assumes that for voiced speech the input consists of a train of impulses whose separation is equal to the pitch period and that for unvoiced speech the input is taken to be white noise. A combination of impulses and noise is used whenever the speech is both voiced and fricative.

Equation (8-1) is the pole-zero vocal tract model. Given a speech sample $w(nT)$ the problem now becomes that of finding the coefficients a_k , b_k , $1 \leq k \leq P$, $1 \leq k \leq Q$. These coefficients completely define the filter $W(z) = W(z)/G(z)$ where $W(z)$ and $G(z)$ are the z-transform of $w(nT)$ and $g(nT)$ respectively. In general, this is a difficult task. It turns out that one must solve a set of $P+Q$ non-linear equations (75) and must resort to using iterative solutions to arrive at some predetermined minimal error. However, it is well known that for spectral voiced sounds the transfer function of the vocal tract consists only of resonances (42). The transfer function thus becomes

$$x(n) = \frac{A}{1 - \sum_{k=1}^p a_k z^{-k}} \quad (8-2)$$

Now this does not in general hold for nasal and fricative speech since the vocal tract generates antiformants for these sounds. Thus, the task is to find p so that (8-2) does as best approximate (8-1) within some error criteria. Makhoul and Rabin [42] argue that the all-pole model represented by (8-2) gives adequate results because the human auditory system is considerably more sensitive to the location of a pole than to the location of a zero. They also point out that the sensitivity of the human ear to the envelope of the speech spectrum is not dependent on the manner with which the spectrum is generated.

The problem now reduces to one of using linear prediction to find the filter coefficients a_k , $1 \leq k \leq p$ and then to find a suitable input, $x(n)$, to implement the filter. We will assume that at a particular instant in time a speech sample $x(n)$, where $x(n)$ is a sampled version of $x(t)$, can be approximated by a linearly weighted sum of past p samples, where p is some integer. Then,

$$x(n) = \sum_{k=1}^p a_k x(n-k)$$

or

$$a_n = \sum_{k=1}^p a_k a_{n-k} \quad (8-3)$$

and define $\hat{y}_n$ by

$$\hat{y}_n = \sum_{k=1}^N a_k x_{n-k}$$

where T is the sampling period, n is the sample number,

$\hat{y}_n$ is the approximation of y_n , and $a_1, a_2, \dots, a_N$ are constants and will be referred to as the linear prediction coefficients.

Let the error between the actual value and the one predicted by (B-3) be given by

$$e_n = y_n - \hat{y}_n = y_n - \sum_{k=1}^N a_k x_{n-k}.$$

Now various error minimization criteria can be used in estimating the a_k 's. The method employed in linear prediction is the least-squares approximation procedure since it leads to a mathematically attractive solution [4]. Defining the total-squared error by E we have

$$E = \sum_n e_n^2 = \sum_n \left[y_n - \sum_{k=1}^N a_k x_{n-k} \right]^2. \quad (\text{B-4})$$

To minimize E we let

$$\partial E / \partial a_L = 0 \quad 1 \leq L \leq N. \quad (\text{B-5})$$

Substituting (B-4) in (B-5) gives

$$\sum_{n=1}^N a_k \left[\sum_{n=k}^N x_{n-k} x_{n-L} \right] - \sum_n x_n x_{n-L} = 0 \quad 1 \leq L \leq N. \quad (\text{B-6})$$

We see that the range of summation in (8-4) has not been specified. This leads us to consider two distinct cases, each one having different assumptions.

In the covariance method the range of n is taken over a finite interval, $0 \leq n \leq N-1$, where N is some integer. Equation (8-4) then reduces to

$$\sum_{k=1}^p a_k R_{1k} = c_{1k} \quad 1 \leq k \leq p$$

where

$$c_{1k} = \sum_{n=0}^{N-1} x_{n+1} x_{n+k}.$$

We thus arrive at a system of p linear equations with p unknowns.

The autocorrelation method makes the assumption that $m < n < M$, and defines r_n^* as follows,

$$r_n^* = \begin{cases} x_n x_{n+1} & 0 \leq n \leq N-1 \\ 0 & \text{otherwise} \end{cases}$$

where x_n is a known window function.

Using these conditions in (8-4), we have

$$\sum_{k=1}^p a_k R_{1k} = R_1 \quad 1 \leq k \leq p$$

where

$$R_1 = \sum_{n=0}^{N-1} x_n x_{n+1}.$$

again we have a system of p linear equations with p unknowns.

The difference between the two methods can be interpreted as one assumes that the signal is nonstationary (covariance method) and the other that the signal is stationary (autocorrelation). Thus, the covariance method reduces to the autocorrelation method whenever the speech signal is stationary. It is possible to give other formulations for these two methods [42], but the ones given here are adequate for our needs.

The effects of radiation at the mouth and nostrils and of glottal flow play an important part in the overall filter design. Parnis [3] reports that the effect of radiation at the mouth and nostrils can be approximated by differentiation. This can be represented by $k_1 \Omega^{-1}$, where k_1 is a constant dependent on the amplitude of the volume flow at the lips and the distance from the lips to the microphone [43]. The overall effect due to glottal shape can be approximated by two poles [44] which leads to an expression of the form $k_2/(1-a_1z^{-1})(1-a_2z^{-1})$ where k_2 is a constant proportional to the intensity of glottal flow, a_1 and a_2 are real poles such that $|a_1|, |a_2| \leq 1$. Thus, the total effect due to radiation and glottal flow is given by the transfer function

$$H(z) = \frac{k_1 k_2 (1-z^{-1})}{(1-a_1 z^{-1})(1-a_2 z^{-1})} \quad (1)$$

And [45] approximates this last expression to an all-pole model

$$W(\omega) = \frac{A_1 A_2}{(1 + i\omega a_1)^{2/3} (1 + i\omega a_2)^{2/3}}$$

and indicates that the approximation is a good one since usually $a_1 \approx 1$.

The linear production model, given by (9-15), takes time effects into consideration [41-44]. The value of p should therefore be chosen appropriately so that it takes into account forward frequencies, gradual flow, and radiation. A good value of p is usually around 15 to 18

APPENDIX C

FOOTNOTES

Source: Authors

EXAMPLES

The examples presented in this section are aimed at illustrating the use of the programs previously listed. The programs are all written in FORTRAN (except for some of the subroutines, which are in assembly language) and were implemented on a Honeywell 6000 computer. The programs are used in conjunction with each other:

In the example below, the program APCOEFFS calculates the filter coefficients and stores them in a disk file. The jitter and shimmer parameters are then derived from the original speech signal by the program FITERR. The program then stores these two parameters in separate disk files. The third program, called APCUTERR, reads in from disk files the filter coefficients, the jitter and shimmer parameters, and the glottal input pulse period. The program then periodically extends the glottal cycle and incorporates the jitter and shimmer parameters. The resultant glottal waveform is filtered through the vocal tract model and the output waveform stored in a file on disk. The final process of incorporating the effect of radiation is done by the program DIFFRAC. The output speech is stored on a specified disk file.

All of the programs are interactive since operator intervention is required to specify input and output data files, and information regarding format, frequencies and bandwidths. The disk program requires the use of a display terminal or an oscilloscope interfaced with the com-

puter for purposes of displaying the frequency responses of the linear prediction filter and the modified filter.

The programs are called from disk files by typing on the terminal the name of the program. Execution is terminated when an R is displayed on the terminal.


```

LOCUTTER=
HOW MANY FACTORIAL COEFFICIENTS IN
THIS FILENAME?
STOP
DOES SLIGHT DATA FILENAME?
STOP
READING POINTS ARE
HOW MANY POINTS IN SLIGHT AREA? PA
DOES DATA FILENAME?
STOP
HOW MANY STRINGS OF SLIGHT AREA? LER
IS AREA PRESENT?
DOES DATA FILENAME?
STOP
IS NUMBER SEQUENCE?
DOES DATA FILENAME?
STOP
DOES SEQUENCE FACTOR? DOES
STOP
?

```

```

EXTENDED
DOES DATA FILENAME?
STOP
DOES AREA?
DOES OF BLOCK? IN
DOES OF A IN STRUCTURE? 47
DOES OF A IN STRUCTURE? 47
DOES OF A IN STRUCTURE? 47
DOES DATA FILENAME?
STOP
DOES DATA FILENAME?
STOP
?

```

APPENDIX B
LEARNER EVALUATION FORM

NAME _____ CLASS/SECTION NO. _____
AGE _____

LOCAL
ADDRESS _____

PHONE _____

TABLE 1

Group 1	Group 2	Group 3
1) W	13) W	15) W
2) W	23) W	25) W
3) W	33) W	35) W
4) W	43) W	45) W
5) W	53) W	55) W
6) W	63) W	65) W
7) W	73) W	75) W
8) W	83) W	85) W
9) W	93) W	95) W
10) W	103) W	105) W
11) W	113) W	115) W
12) W	123) W	125) W
13) W	133) W	135) W
14) W	143) W	145) W
15) W	153) W	155) W
16) W	163) W	165) W
17) W	173) W	175) W
18) W	183) W	185) W
19) W	193) W	195) W
20) W	203) W	205) W
21) W	213) W	215) W
22) W	223) W	225) W
23) W	233) W	235) W
24) W	243) W	245) W
25) W	253) W	255) W
26) W	263) W	265) W
27) W	273) W	275) W
28) W	283) W	285) W
29) W	293) W	295) W
30) W	303) W	305) W
31) W	313) W	315) W
32) W	323) W	325) W
33) W	333) W	335) W
34) W	343) W	345) W
35) W	353) W	355) W
36) W	363) W	365) W
37) W	373) W	375) W
38) W	383) W	385) W
39) W	393) W	395) W
40) W	403) W	405) W
41) W	413) W	415) W
42) W	423) W	425) W
43) W	433) W	435) W
44) W	443) W	445) W
45) W	453) W	455) W
46) W	463) W	465) W
47) W	473) W	475) W
48) W	483) W	485) W
49) W	493) W	495) W
50) W	503) W	505) W
51) W	513) W	515) W
52) W	523) W	525) W
53) W	533) W	535) W
54) W	543) W	545) W
55) W	553) W	555) W
56) W	563) W	565) W
57) W	573) W	575) W
58) W	583) W	585) W
59) W	593) W	595) W
60) W	603) W	605) W
61) W	613) W	615) W
62) W	623) W	625) W
63) W	633) W	635) W
64) W	643) W	645) W
65) W	653) W	655) W
66) W	663) W	665) W
67) W	673) W	675) W
68) W	683) W	685) W
69) W	693) W	695) W
70) W	703) W	705) W
71) W	713) W	715) W
72) W	723) W	725) W
73) W	733) W	735) W
74) W	743) W	745) W
75) W	753) W	755) W
76) W	763) W	765) W
77) W	773) W	775) W
78) W	783) W	785) W
79) W	793) W	795) W
80) W	803) W	805) W
81) W	813) W	815) W
82) W	823) W	825) W
83) W	833) W	835) W
84) W	843) W	845) W
85) W	853) W	855) W
86) W	863) W	865) W
87) W	873) W	875) W
88) W	883) W	885) W
89) W	893) W	895) W
90) W	903) W	905) W
91) W	913) W	915) W
92) W	923) W	925) W
93) W	933) W	935) W
94) W	943) W	945) W
95) W	953) W	955) W
96) W	963) W	965) W
97) W	973) W	975) W
98) W	983) W	985) W
99) W	993) W	995) W
100) W	1003) W	1005) W
101) W	1013) W	1015) W
102) W	1023) W	1025) W
103) W	1033) W	1035) W
104) W	1043) W	1045) W
105) W	1053) W	1055) W
106) W	1063) W	1065) W
107) W	1073) W	1075) W
108) W	1083) W	1085) W
109) W	1093) W	1095) W
110) W	1103) W	1105) W
111) W	1113) W	1115) W
112) W	1123) W	1125) W
113) W	1133) W	1135) W
114) W	1143) W	1145) W
115) W	1153) W	1155) W
116) W	1163) W	1165) W
117) W	1173) W	1175) W
118) W	1183) W	1185) W
119) W	1193) W	1195) W
120) W	1203) W	1205) W
121) W	1213) W	1215) W
122) W	1223) W	1225) W
123) W	1233) W	1235) W
124) W	1243) W	1245) W
125) W	1253) W	1255) W
126) W	1263) W	1265) W
127) W	1273) W	1275) W
128) W	1283) W	1285) W
129) W	1293) W	1295) W
130) W	1303) W	1305) W
131) W	1313) W	1315) W
132) W	1323) W	1325) W
133) W	1333) W	1335) W
134) W	1343) W	1345) W
135) W	1353) W	1355) W
136) W	1363) W	1365) W
137) W	1373) W	1375) W
138) W	1383) W	1385) W
139) W	1393) W	1395) W
140) W	1403) W	1405) W
141) W	1413) W	1415) W
142) W	1423) W	1425) W
143) W	1433) W	1435) W
144) W	1443) W	1445) W
145) W	1453) W	1455) W
146) W	1463) W	1465) W
147) W	1473) W	1475) W
148) W	1483) W	1485) W
149) W	1493) W	1495) W
150) W	1503) W	1505) W
151) W	1513) W	1515) W
152) W	1523) W	1525) W
153) W	1533) W	1535) W
154) W	1543) W	1545) W
155) W	1553) W	1555) W
156) W	1563) W	1565) W
157) W	1573) W	1575) W
158) W	1583) W	1585) W
159) W	1593) W	1595) W
160) W	1603) W	1605) W
161) W	1613) W	1615) W
162) W	1623) W	1625) W
163) W	1633) W	1635) W
164) W	1643) W	1645) W
165) W	1653) W	1655) W
166) W	1663) W	1665) W
167) W	1673) W	1675) W
168) W	1683) W	1685) W
169) W	1693) W	1695) W
170) W	1703) W	1705) W
171) W	1713) W	1715) W
172) W	1723) W	1725) W
173) W	1733) W	1735) W
174) W	1743) W	1745) W
175) W	1753) W	1755) W
176) W	1763) W	1765) W
177) W	1773) W	1775) W
178) W	1783) W	1785) W
179) W	1793) W	1795) W
180) W	1803) W	1805) W
181) W	1813) W	1815) W
182) W	1823) W	1825) W
183) W	1833) W	1835) W
184) W	1843) W	1845) W
185) W	1853) W	1855) W
186) W	1863) W	1865) W
187) W	1873) W	1875) W
188) W	1883) W	1885) W
189) W	1893) W	1895) W
190) W	1903) W	1905) W
191) W	1913) W	1915) W
192) W	1923) W	1925) W
193) W	1933) W	1935) W
194) W	1943) W	1945) W
195) W	1953) W	1955) W
196) W	1963) W	1965) W
197) W	1973) W	1975) W
198) W	1983) W	1985) W
199) W	1993) W	1995) W
200) W	2003) W	2005) W
201) W	2013) W	2015) W
202) W	2023) W	2025) W
203) W	2033) W	2035) W
204) W	2043) W	2045) W
205) W	2053) W	2055) W
206) W	2063) W	2065) W
207) W	2073) W	2075) W
208) W	2083) W	2085) W
209) W	2093) W	2095) W
210) W	2103) W	2105) W
211) W	2113) W	2115) W
212) W	2123) W	2125) W
213) W	2133) W	2135) W
214) W	2143) W	2145) W
215) W	2153) W	2155) W
216) W	2163) W	2165) W
217) W	2173) W	2175) W
218) W	2183) W	2185) W
219) W	2193) W	2195) W
220) W	2203) W	2205) W
221) W	2213) W	2215) W
222) W	2223) W	2225) W
223) W	2233) W	2235) W
224) W	2243) W	2245) W
225) W	2253) W	2255) W
226) W	2263) W	2265) W
227) W	2273) W	2275) W
228) W	2283) W	2285) W
229) W	2293) W	2295) W
230) W	2303) W	2305) W
231) W	2313) W	2315) W
232) W	2323) W	2325) W
233) W	2333) W	2335) W
234) W	2343) W	2345) W
235) W	2353) W	2355) W
236) W	2363) W	2365) W
237) W	2373) W	2375) W
238) W	2383) W	2385) W
239) W	2393) W	2395) W
240) W	2403) W	2405) W
241) W	2413) W	2415) W
242) W	2423) W	2425) W
243) W	2433) W	2435) W
244) W	2443) W	2445) W
245) W	2453) W	2455) W
246) W	2463) W	2465) W
247) W	2473) W	2475) W
248) W	2483) W	2485) W
249) W	2493) W	2495) W
250) W	2503) W	2505) W
251) W	2513) W	2515) W
252) W	2523) W	2525) W
253) W	2533) W	2535) W
254) W	2543) W	2545) W
255) W	2553) W	2555) W
256) W	2563) W	2565) W
257) W	2573) W	2575) W
258) W	2583) W	2585) W
259) W	2593) W	2595) W
260) W	2603) W	2605) W
261) W	2613) W	2615) W
262) W	2623) W	2625) W
263) W	2633) W	2635) W
264) W	2643) W	2645) W
265) W	2653) W	2655) W
266) W	2663) W	2665) W
267) W	2673) W	2675) W
268) W	2683) W	2685) W
269) W	2693) W	2695) W
270) W	2703) W	2705) W
271) W	2713) W	2715) W
272) W	2723) W	2725) W
273) W	2733) W	2735) W
274) W	2743) W	2745) W
275) W	2753) W	2755) W
276) W	2763) W	2765) W
277) W	2773) W	2775) W
278) W	2783) W	2785) W
279) W	2793) W	2795) W
280) W	2803) W	2805) W
281) W	2813) W	2815) W
282) W	2823) W	2825) W
283) W	2833) W	2835) W
284) W	2843) W	2845) W
285) W	2853) W	2855) W
286) W	2863) W	2865) W
287) W	2873) W	2875) W
288) W	2883) W	2885) W
289) W	2893) W	2895) W
290) W	2903) W	2905) W
291) W	2913) W	2915) W
292) W	2923) W	2925) W
293) W	2933) W	2935) W
294) W	2943) W	2945) W
295) W	2953) W	2955) W
296) W	2963) W	2965) W
297) W	2973) W	2975) W
298) W	2983) W	2985) W
299) W	2993) W	2995) W
300) W	3003) W	3005) W
301) W	3013) W	3015) W
302) W	3023) W	3025) W
303) W	3033) W	3035) W
304) W	3043) W	3045) W
305) W	3053) W	3055) W
306) W	3063) W	3065) W
307) W	3073) W	3075) W
308) W	3083) W	3085) W
309) W	3093) W	3095) W
310) W	3103) W	3105) W
311) W	3113) W	3115) W
312) W	3123) W	3125) W
313) W	3133) W	3135) W
314) W	3143) W	3145) W
315) W	3153) W	3155) W
316) W	3163) W	3165) W
317) W	3173) W	3175) W
318) W	3183) W	3185) W
319) W	3193) W	3195) W
320) W	3203) W	3205) W
321) W	3213) W	3215) W
322) W	3223) W	3225) W
323) W	3233) W	3235) W
324) W	3243) W	3245) W
325) W	3253) W	3255) W
326) W	3263) W	3265) W
327) W	3273) W	3275) W
328) W	3283) W	3285) W
329) W	3293) W	3295) W
330) W	3303) W	3305) W
331) W	3313) W	3315) W
332) W	3323) W	3325) W
333) W	3333) W	3335) W
334) W	3343) W	3345) W
335) W	3353) W	3355) W
336) W	3363) W	3365) W
337) W	3373) W	3375) W
338) W	3383) W	3385) W
339) W	3393) W	3395) W
340) W	3403) W	3405) W

TABLE 1 - continued

Group 4	Group 5
15 W 0	14 W 0
21 W 0	20 W 0
30 W 0	30 W 0
40 W 0	40 W 0
51 W 0	50 W 0
61 W 0	60 W 0
71 W 0	70 W 0
81 W 0	80 W 0
91 W 0	90 W 0
101 W 0	100 W 0
111 W 0	110 W 0
120 W 0	120 W 0
130 W 0	130 W 0
140 W 0	140 W 0
150 W 0	150 W 0
160 W 0	160 W 0
170 W 0	170 W 0
181 W 0	180 W 0

TABLE 21

Group 1	Group 2	Group 3	Group 4
1) a b	11) a b	11) a b	11) a b
2) a b	21) a b	21) a b	21) a b
30) a b	31) a b	31) a b	31) a b
42) a b	41) a b	41) a b	41) a b
50) a b	51) a b	51) a b	51) a b
62) a b	61) a b	61) a b	61) a b
72) a b	71) a b	71) a b	71) a b
81) a b	81) a b	81) a b	81) a b
81) a b	91) a b	81) a b	10) a b
101) a b	101) a b	101) a b	101) a b
111) a b	111) a b	111) a b	111) a b
121) a b	121) a b	121) a b	121) a b

TABLE 11: $\mathcal{C}$ (continued)

Group 5	Group 6	Group 7	Group 8
1) a, b	31) a, b	1) a, b	31) a, b
20) a, b	32) a, b	20) a, b	32) a, b
31) a, b	39) a, b	31) a, b	39) a, b
41) a, b	49) a, b	41) a, b	49) a, b
51) a, b	59) a, b	51) a, b	59) a, b
61) a, b	69) a, b	61) a, b	69) a, b
79) a, b	71) a, b	79) a, b	71) a, b
89) a, b	81) a, b	89) a, b	81) a, b
91) a, b	99) a, b	91) a, b	99) a, b
141) a, b	149) a, b	141) a, b	149) a, b
151) a, b	159) a, b	151) a, b	159) a, b
199) a, b	121) a, b	199) a, b	121) a, b

TABLE C.11

Group 1

1) same _____
different _____

2) same _____
different _____

3) same _____
different _____

4) same _____
different _____

5) same _____
different _____

6) same _____
different _____

7) same _____
different _____

8) same _____
different _____

9) same _____
different _____

10) same _____
different _____

11) same _____
different _____

12) same _____
different _____

Group 2

1) same _____
different _____

2) same _____
different _____

3) same _____
different _____

4) same _____
different _____

5) same _____
different _____

6) same _____
different _____

7) same _____
different _____

8) same _____
different _____

9) same _____
different _____

10) same _____
different _____

11) same _____
different _____

12) same _____
different _____

APPENDIX C
DATA FOR SUBJECTS MB AND JB

The data for subjects MB and JB are presented in this appendix. The tables below correspond to Table 3.1 of subject JB. The figures presented here for the two subjects follow those of subject JB, that is, Figures 3-1 to 3-5 and 3-6 to 3-18 correspond to Figures 3.1 to 3.5 in Chapter 3.

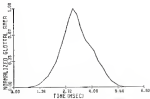
The analysis of the data for these two subjects is similar to that of subject JB. An apparent difference is that for these two subjects the spectrum of the synthetic speech, the frequencies below half the carrier's sampling frequency, does not coincide with the spectrum of the real speech as well as that for subject JB. This can probably be attributed to the high fundamental frequency of subjects MB and JB and the approximations to the measurement of the glottal area made by the operator for these two subjects.

Table B-1. Foremost locations and bandwidths determined from the linear prediction and modified filter models for subject BB.

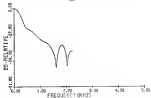
	F_1	F_2	F_3	F_4
Foremost Location (kHz)	.48	1.44	3.32	3.87
Linear Prediction Bandwidth (kHz)	218	193	184	219
Modified Filter Bandwidth (kHz)	82	78	115	175

Table B-2. Foremost locations and bandwidths as derived from the linear prediction and modified filter models for subject AB.

	F_1	F_2	F_3	F_4
Foremost Location (kHz)	0.5	1.43	3.18	3.83
Linear Prediction Bandwidth (kHz)	148	88	194	188
Modified Filter Bandwidth (kHz)	43	78	115	175

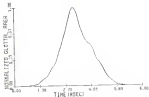


(a)

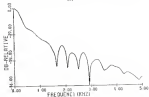


(b)

Figure 8-3. a) First period of glassy state waveform of subject 88- to FFT of above. A rectangular window was used.

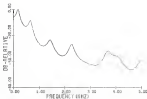


(a)

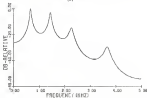


(b)

Figure B-1. (a) Pitch period of interpolated glacial area waveform of subject 28. (b) FFT of above. A rectangular window was used.

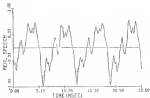


(a)

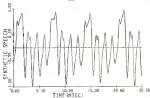


(b)

FIGURE 8-2 Frequency response of a) linear prediction filter at order $p = 16$, and b) modified vocal tract filter at order $p = 0$. As if no speech formed all subject HR was used for the analysis.

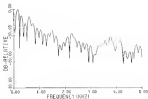


(a)

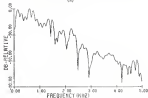


(b)

Figure B-3 a) Natural speech of subject BB.
 b) Synthetic speech of subject BB.



(a)



(b)

FIGURE 8-4 Spectrum of a) natural speech, and b) synthetic speech of subject 78. A harmonic window was used.

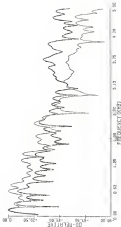
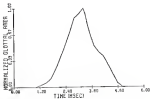
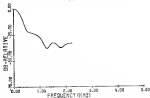


Figure 2-3 Spectra of real (dark shaded line) and simulated (lighter shaded line) speech waveforms of subject 25. A smoothing window was used.

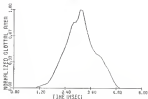


(a)

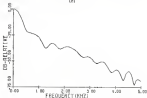


(b)

Figure D-8 (a) P100 period of digital test waveform of subject DB. (b) FFT of above. A rectangular window was used.

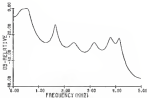


(a)

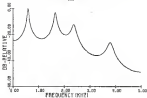


(b)

Figure 2-1 (a) Pitch period of unprocessed glottal area waveform of subject 28 (b) rfr of curve; A rectangular window was used

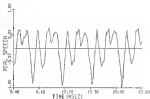


(a)

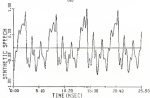


(b)

Figure B-3 Frequency response of a) linear predictive filter of order $p = 16$, and b) modified vocal tract filter of order $p = 4$. An 18 ms speech record of subject 18 was used for the analysis.

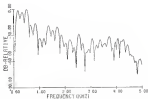


(a)

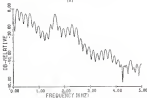


(b)

Figure 2-8 a) Natural speech of subject 28
 b) Synthetic speech of subject 28.



(a)



(b)

Figure 5-8. Spectrum of (a) natural speech, and (b) synthetic speech of subject 26. A Hanning window was used.

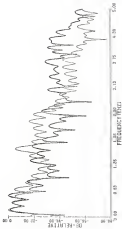


Figure 2-18 Spectrum of acid. (Dark shaded line) and spectrum (light shaded line) showing variations of analogs. (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, aa, ab, ac, ad, ae, af, ag, ah, ai, aj, ak, al, am, an, ao, ap, aq, ar, as, at, au, av, aw, ax, ay, az, ba, bb, bc, bd, be, bf, bg, bh, bi, bj, bk, bl, bm, bn, bo, bp, bq, br, bs, bt, bu, bv, bw, bx, by, bz, ca, cb, cc, cd, ce, cf, cg, ch, ci, cj, ck, cl, cm, cn, co, cp, cq, cr, cs, ct, cu, cv, cw, cx, cy, cz, da, db, dc, dd, de, df, dg, dh, di, dj, dk, dl, dm, dn, do, dp, dq, dr, ds, dt, du, dv, dw, dx, dy, dz, ea, eb, ec, ed, ee, ef, eg, eh, ei, ej, ek, el, em, en, eo, ep, eq, er, es, et, eu, ev, ew, ex, ey, ez, fa, fb, fc, fd, fe, ff, fg, fh, fi, fj, fk, fl, fm, fn, fo, fp, fq, fr, fs, ft, fu, fv, fw, fx, fy, fz, ga, gb, gc, gd, ge, gf, gg, gh, gi, gj, gk, gl, gm, gn, go, gp, gq, gr, gs, gt, gu, gv, gw, gx, gy, gz, ha, hb, hc, hd, he, hf, hg, hh, hi, hj, hk, hl, hm, hn, ho, hp, hq, hr, hs, ht, hu, hv, hw, hx, hy, hz, ia, ib, ic, id, ie, if, ig, ih, ii, ij, ik, il, im, in, io, ip, iq, ir, is, it, iu, iv, iw, ix, iy, iz, ja, jb, jc, jd, je, jf, jg, jh, ji, jj, jk, jl, jm, jn, jo, jp, jq, jr, js, jt, ju, jv, jw, jx, jy, jz, ka, kb, kc, kd, ke, kf, kg, kh, ki, kj, kk, kl, km, kn, ko, kp, kq, kr, ks, kt, ku, kv, kw, kx, ky, kz, la, lb, lc, ld, le, lf, lg, lh, li, lj, lk, ll, lm, ln, lo, lp, lq, lr, ls, lt, lu, lv, lw, lx, ly, lz, ma, mb, mc, md, me, mf, mg, mh, mi, mj, mk, ml, mm, mn, mo, mp, mq, mr, ms, mt, mu, mv, mw, mx, my, mz, na, nb, nc, nd, ne, nf, ng, nh, ni, nj, nk, nl, nm, nn, no, np, nq, nr, ns, nt, nu, nv, nw, nx, ny, nz, oa, ob, oc, od, oe, of, og, oh, oi, oj, ok, ol, om, on, oo, op, oq, or, os, ot, ou, ov, ow, ox, oy, oz, pa, pb, pc, pd, pe, pf, pg, ph, pi, pj, pk, pl, pm, pn, po, pp, pq, pr, ps, pt, pu, pv, pw, px, py, pz, qa, qb, qc, qd, qe, qf, qg, qh, qi, qj, qk, ql, qm, qn, qo, qp, qq, qr, qs, qt, qu, qv, qw, qx, qy, qz, ra, rb, rc, rd, re, rf, rg, rh, ri, rj, rk, rl, rm, rn, ro, rp, rq, rr, rs, rt, ru, rv, rw, rx, ry, rz, sa, sb, sc, sd, se, sf, sg, sh, si, sj, sk, sl, sm, sn, so, sp, sq, sr, ss, st, su, sv, sw, sx, sy, sz, ta, tb, tc, td, te, tf, tg, th, ti, tj, tk, tl, tm, tn, to, tp, tq, tr, ts, tt, tu, tv, tw, tx, ty, tz, ua, ub, uc, ud, ue, uf, ug, uh, ui, uj, uk, ul, um, un, uo, up, uq, ur, us, ut, uu, uv, uw, ux, uy, uz, va, vb, vc, vd, ve, vf, vg, vh, vi, vj, vk, vl, vm, vn, vo, vp, vq, vr, vs, vt, vu, vv, vw, vx, vy, vz, wa, wb, wc, wd, we, wf, wg, wh, wi, wj, wk, wl, wm, wn, wo, wp, wq, wr, ws, wt, wu, wv, ww, wx, wy, wz, xa, xb, xc, xd, xe, xf, xg, xh, xi, xj, xk, xl, xm, xn, xo, xp, xq, xr, xs, xt, xu, xv, xw, xx, xy, xz, ya, yb, yc, yd, ye, yf, yg, yh, yi, yj, yk, yl, ym, yn, yo, yp, yq, yr, ys, yt, yu, yv, yw, yx, yy, yz, za, zb, zc, zd, ze, zf, zg, zh, zi, zj, zk, zl, zm, zn, zo, zp, zq, zr, zs, zt, zu, zv, zw, zx, zy, zz).

REFERENCES

- [1] F.C. Flanagan, Speech Analysis Synthesis and Prediction, 2nd Edition. New York: Springer-Verlag, 1972.
- [2] J.D. Markel and A.H. Gray, Jr., Linear Prediction of Speech. New York: Springer-Verlag, 1976.
- [3] G. Sullivan and J. Lindqvist, "Spectrotemporal measurements of vocal tract characteristics," J. Acoust. Soc. Am., vol. 49, pp. 1412-1420, Feb. 1971.
- [4] C.R. Rabiner and B. Gold, Theory and Application of Digital Signal Processing. New Jersey: Prentice-Hall, 1975.
- [5] C.F. Moore, F.D. Miller, and E. von Leden, "Using high speed photography in laryngeal physiology," J. Speech Hear. Dis., vol. 27, pp. 162-171, May 1962.
- [6] R.J. Jordan, Correlation of Glottal Area Function Using Three Methods as Measured Through Measurement of Wave-Height-Speed Relationships and Fluid-Flow-Field Characteristics, Ph.D. Dissertation, University of Florida, 1973.
- [7] R.B. Stidebrand, Vibratory Patterns of the Human Vocal Cords During Vocalization as Measured by Pharyngography, Ph.D. Dissertation, University of Florida, 1974.
- [8] C.M. Wells, Physical Acoustical Relationships, 2nd Edition. New York: McGraw-Hill, 1968.
- [9] A.E. Sussman, "Effect of signal pulse shape on the gliding of natural vowels," J. Acoust. Soc. Am., vol. 49, pp. 1471-1472, Feb. 1971.
- [10] C.M. Wells, "The influence of glottal waveform on the characteristics of speech from a parallel formed system," IEEE Trans. Audio Electroacoustics, vol. AP-21, pp. 249-253, June 1973.
- [11] J.F. Moore, Personal communication, July 24, 1978.
- [12] J.L. Flanagan, "Analysis of intraglottal pressure during phonation," J. Speech Res. Soc., vol. 2, pp. 148-173, June 1959.

- [123] J.A. Flanagan and R.E.S. Meddisart, "Source-filter synthesis in the vocal tract," J. Acoust. Soc. Am., vol. 34, p. 1864(18), Oct. 1962.
- [124] J.L. Flanagan and L.L. Lindgren, "Self-excited oscillations for vocal-tract synthesis," IEEE Trans. Audio Electroacoust., vol. 10-14, pp. 10-14, Mar. 1962.
- [125] J.L. Flanagan, "Use of an interactive laboratory computer to study an acoustic oscillator model of the vocal cords," IEEE Trans. Audio Electroacoust., vol. 10-17, pp. 2-4, Mar. 1969.
- [126] J.L. Flanagan and L. Cherry, "Excitation of vocal-tract synthesizers," J. Acoust. Soc. Am., vol. 43, pp. 714-719, June 1969.
- [127] R. Ishizaka and J.L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," IEEE System Tech. J., vol. 80, pp. 1213-1218, July 1972.
- [128] J.M. Van Den Burg, J.T. Hartman, and P. Donkersloot, Jr., "On the air resistance and the Bernoulli effect of the human larynx," J. Acoust. Soc. Am., vol. 38, pp. 416-421, May 1967.
- [129] J.L. Flanagan, "The synthesis of speech," Scientific Amer., vol. 124, pp. 14-24, Feb. 1972.
- [130] J.L. Flanagan, "Voices of men and machines," J. Acoust. Soc. Am., vol. 51, pp. 1379-1387, May 1972.
- [131] J.B. Whittle, Speech Synthesis. London: Mills and Boon, 1971.
- [132] J.D. Stewart, "An electrical analogue of the vocal organs," Science, vol. 110, pp. 341-352, Sept. 1923.
- [133] R. Dudley, R. Egan, and J. Mathews, "A synthetic speaker," J. Franklin Inst., vol. 217, pp. 719-744, June 1935.
- [134] R. Dudley, "The carrier waves of speech," IEEE System Tech. J., vol. 17, pp. 493-503, Oct. 1948.
- [135] R.L. Egan and E.L. Hansen, "Artificial speech in phonetic and communications," J. Speech Res. Soc., vol. 1, pp. 13-15, Mar. 1958.
- [136] C.M. Harris, "A speech synthesizer," J. Acoust. Soc. Am., vol. 35, pp. 878-879, Sept. 1963.
- [137] C.M. Harris, "A study of the building blocks in speech," J. Acoust. Soc. Am., vol. 35, pp. 842-849, Sept. 1963.

- [18] R.E. Dunn, "The calculation of vocal resonances and an electrical vocal tract," J. Acoust. Soc. Am., vol. 32, pp. 148-151, Nov. 1960.
- [19] R.E. Stevens, S. Banerjee, and S. Fant, "An electrical analog of the vocal tract," J. Acoust. Soc. Am., vol. 35, pp. 714-743, July 1963.
- [20] J.L. Flanagan, "Notes on the design of 'Formant-singing' speech synthesizers," J. Acoust. Soc. Am., vol. 39, pp. 209-218, Feb. 1967.
- [21] M. Levinson, "The synthesis of speech from signals which have a low information rate," pp. 448-460 in: Communication Theory, London: Butterworths Sci. Publ., 1961.
- [22] J.M. Bahner, L.W. Jackson, R.M. Schaffer and C.W. Cohen, "A hardware realization of a digital formant speech synthesizer," IEEE Trans. Comput., vol. COM-19, Dec., pp. 1414-1420, 1971.
- [23] J. Litvinovskaya, "The DVE III speech synthesizer," IEEE Trans. Audio Electroacoustics, vol. AU-14, pp. 137-148, Mar. 1966.
- [24] R.E. Furdock, "SPUR - a terminal analog speech synthesizer," Acoustics, vol. 32, no. 8, pp. 197-202, 1960.
- [25] J.M. Bahner, "A model for synthesizing speech by voice," IEEE Trans. Audio Electroacoustics, AU-17, pp. 7-13, Dec. 1969.
- [26] J.L. Flanagan, C.W. Cohen, L.M. Bahner, R.M. Schaffer, and S. Umeda, "Synthesizing vowels for computers," IEEE Spectrum, vol. 7, pp. 32-43, Oct. 1970.
- [27] A.V. Oppenheim and R.M. Schaffer, "Homomorphic analysis of speech," IEEE Trans. Audio Electroacoustics, vol. AU-14, pp. 210-224, June 1966.
- [28] A.V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," J. Acoust. Soc. Am., vol. 43, pp. 454-465, Feb. 1968.
- [29] J.B. Markel, "Digital general filtering - a new tool for formant trajectory estimation," IEEE Trans. Audio Electroacoustics, vol. AU-20, pp. 118-127, June 1972.
- [30] F. Itakura and S. Saito, "Analysis-synthesis telephony based on the minimum likelihood method," Trans. Electr. Inf. Comm. Acoust., Paper 6-5-5, pp. 617-68, Aug. 1968.
- [31] R.L. Atal and S.L. Hancock, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., vol. 50, pp. 637-655, Feb. 1973.

- [42] J. J. Mathew and J. J. Melf, "Formant prediction and the spectral analysis of speech," IEEE AG-303040, Bell Research and Research, Inc., Cambridge, Mass., 196 Sep. 1964, Aug. 1972.
- [43] C. D. Bell, R. Fujisaki, J. V. Serna, S. S. Stevens, and A. S. House, "Synthesis of speech spectra by analysis-by-synthesis techniques," J. Acoust. Soc. Am., vol. 49, pp. 1723-1729, Dec. 1971.
- [44] J. J. Mathew, "Spectral linear prediction: properties and applications," IEEE Trans. Acoust. Speech, Signal Processing, vol. ASSP-25, pp. 381-394, June 1975.
- [45] J. L. Flanagan, E. Ishizaka, and R. L. Shipley, "Formants of speech from a dynamic model of the vocal cords and vocal tract," Bell System Tech. J., vol. 54, pp. 819-828, Mar. 1975.
- [46] J. L. Flanagan and E. Ishizaka, "Automatic generation of sinusoidal excitation in a vocal cordless tract speech synthesizer," IEEE Trans. Acoust. Speech, Signal Processing, vol. ASSP-25, pp. 183-179, April 1974.
- [47] J. J. Mathew, "Spectral analysis of speech by linear prediction," IEEE Trans. Audio Electroacoust., vol. AP-22, pp. 160-168, June 1974.
- [48] R. W. Mathew, "Laryngeal noise synthesis of voiced and unvoiced auditory processing of formants," IEEE Trans., vol. 44, no. 3, pp. 99-108, 1970.
- [49] E. F. Coleman, "Effect of waveform changes upon roughness perception," IEEE Trans., vol. 21, no. 3, pp. 314-322, 1970.
- [50] J. L. Flanagan, "Random measurements of voiced excitation from the mouth," J. Acoust. Soc. Am., vol. 32, pp. 3413-3418, Dec. 1962.
- [51] D. W. Green and J. V. Serna, Signal Detection Theory and Psychophysics. New York: Wiley, 1964.
- [52] J. V. Serna, Signal Detection and Recognition by Human Observers. New York: Wiley, 1974.
- [53] E. Tieska, E. von Leden, and E. F. Moore, "Laryngeal vibrational measurements of the glottic wave. Part I: The normal vibratory cycle," Arch. Otolaryngol., vol. 88, pp. 3-19, July 1974.
- [54] E. Tieska, E. von Leden, and E. F. Moore, "Laryngeal vibrational measurements of the glottic wave. Part II: Populating a vibratory cycle," Arch. Otolaryngol., vol. 89, pp. 418-441, April 1975.

- [15] S. von Lohse, G.P. Moore, and R. Linsley, "Laryngeal vibrations: measurements of the glottis wave. Part III. The pathological larynx," Arch. Otolaryngol., vol. 72, pp. 14-20, Jan. 1960.
- [16] L. Hedstrom, A semi-automated system for GOATED SPEECH MEASUREMENTS FROM HIGH SPEED LARYNGEAL FILM, Master's Degree Thesis, University of Florida, 1975.
- [17] D.G. Childers, A. Payne, G.P. Moore, and M. Medal-Durfa, "Nonlinear-coded measurements of laryngeal vibration patterns from high speed film," Proc. IEEE Acoustics Conf., Paper 258, pp. 1-3, April 1974.
- [18] D.G. Childers, A. Payne, G.P. Moore and M. Medal-Durfa, "Automation of the measurement of laryngeal vibration patterns from high speed film," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, April 13-14, pp. 73-84, IEEE Catalog No. T601047-1 Apr., 1974.
- [19] J.L. Bacon, "High-speed photography in speech research," J. Speech Hear. Res., vol. 10, pp. 748-774, Dec. 1967.
- [20] E.B. Napier and P. Soikin, "A data processing system for frame by frame film analysis," Film Review, vol. 24, pp. 109-121, 1973.
- [21] J.L. Flanagan, "Some properties of the glottal sound source," J. Speech Hearing Res., vol. 1, pp. 70-114, June 1968.
- [22] R.B. Krimpholtz, Personal communication, July 11, 1974.
- [23] R. Garcia, M. Strupat, and A. Garza, "A noise source taking account of coupling with the supraglottal cavities," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, April 13-14, pp. 47-50, IEEE Catalog No. T601047-1 Apr., 1974.
- [24] M. Carlsöö, J.-G. Courben, J. Sene and J.P. Loeve, "A hardware vocal source simulator," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, April 13-14, pp. 21-24, IEEE Catalog No. T601047-1 Apr., 1974.
- [25] R. Kane, Investigation of Pharyngeal Voice Source Mechanisms, Ph.D. Dissertation, Massachusetts Institute of Technology, 1973.
- [26] T.B. Crymka, "Model of larynx activity during phonation," M.S.T. Research Lab. of Microphones, Cambridge, Engl., Prep. Rep. 78, July 12, 1962.
- [27] P. Marquardt, "An extension of Flanagan's model of vocal-oid excitation," J. Acoust. Soc. Am., vol. 50, pp. 1208-1218, Oct. 1971.

- [10] G.E. Dedering, "Two-part model of the vocal cords," J. Acoust. Soc. Am., vol. 48, p. 1184-85, July 1970.
- [11] E. Ishizuka and M. Nishikawa, "What makes the vocal cords vibrate," JASA, 51B, 191, COMET, Acoust., pt. 11, pp. 91-13, Aug. 1971.
- [12] L.S. Tasse, The Human Vocal Cords: A Mathematical Model, Ph.D. Dissertation, Brigham Young University, 1971.
- [13] J.L. McDonald, "Vocal production: a tutorial review," JASA, 1972, vol. 43, pp. 543-589, April 1972.
- [14] J.L. McDonald and A.E. Gray, "An impedance/admittance approach as applied to speech analysis," JASA, 1972, Acoust. Electroacoust., vol. 48-11, pp. 61-75, April 1972.

BIOGRAPHICAL SKETCH

Born in New York City on October 20, 1917, Mariano Rafael Merin lived there while completing his secondary education at Bishop Neenan High School.

In 1934 he moved to Puerto Rico to further his education at the University of Puerto Rico, Mayaguez Campus. During the years 1934 to 1941 he finished his Bachelor of Science degree, with high honors, in both Electrical Engineering and Mathematics, and from 1941 to 1943 held a graduate teaching assignment in the University of Puerto Rico's Department of Mathematics.

He accepted, in 1943, a teaching position as an assistant, in the Department of Electrical Engineering, at the same university. He taught from 1943 to 1944 and simultaneously worked on his Master's thesis in the field of Mathematics. Upon completion of his Master's degree in July, 1944, he was granted a leave of absence from the University of Puerto Rico to pursue doctoral studies in Electrical Engineering, with a concentration in High Engineering, at the University of Florida.

He is a member of IEEE, Sigma Xi, CMAA, and is a licensed engineer in Puerto Rico.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.


DONALD T. GLEDHILL, Chairman
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.


R. EARL MOORE
Professor of Speech

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.


ARNOLD F. PIQUE
Associate Professor of Electrical
Engineering and Speech

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.


LOUIS W. CONNER
Associate Professor of Electrical
Engineering

This dissertation was submitted to the Graduate Faculty of
the College of Engineering and to the Graduate Council, and
was accepted as partial fulfillment of the requirements for
the degree of Doctor of Philosophy.

December, 1954



Dean, College of Engineering

Dean, Graduate School